8-2016

Graphical methods in RNA structure matching

Jiajie Huang
Purdue University

Graduate School Form
30 Updated 12/26/2015

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Jiajie Huang

Entitled

GRAPHICAL METHODS IN RNA STRUCTURE MATCHING

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Michael Gribskov
Chair

Cynthia Stauffacher

Daisuke Kihara

Dan Goldwasser

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy of Integrity in Research" and the use of copyright material.

Approved by Major Professor(s): Michael Gribskov

Approved by: Christine Hrycyna 5/5/2016

Head of the Departmental Graduate Program Date

GRAPHICAL METHODS IN RNA STRUCTURE MATCHING

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Jiajie Huang

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

August 2016

Purdue University

West Lafayette, Indiana

Dedicated to the memories of my grandfather.

ACKNOWLEDGEMENTS

I am deeply grateful to my Ph.D. supervisor, Dr. Michael Gribskov, who led me through the adventure of science, enlightened me with smart ideas, and supported me with selfless mind. I also want show my gratitude to my committee members, Dr. Cynthia Stauffacher, Dr. Daisuke Kihara, and Dr. Dan Goldwasser, and former committee member, Dr. Alan Qi, for their selfless help and precious advice in research.

I would also like to thank my labmates for so many years of accompany and support. I am also grateful that I've made many friends at Purdue who care about me and make here my second home.

Last, and most importantly, I want to thank my family, especially my parents, for their unconditional love and support. Without them my dream never comes true.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

LIST OF ABBREVIATIONS

DFS	Depth First Search
MFE	Minimum Free Energy
ncRNA	non-coding RNA
PDB	Protein Data Bank
RAG	RNA As Graphs
RNA	Ribonucleic acid
STRAND	The RNA secondary STRucture and statistical Analysis Database
XIOS	eXclusive, Included, Overlap and Serial

ABSTRACT

Huang, Jiajie. Ph.D., Purdue University, August 2016. Graphical Methods in RNA Structure Matching. Major Professor: Michael Gribskov.

Eukaryotic genomes are pervasively transcribed (1); almost every base can be found in an RNA transcript. This is a surprising observation since most of the genome does not encode proteins. This RNA must serve an important regulatory function – important because producing non-coding RNA is an energy intensive process, and in the absence of strong selection one would expect it to disappear.

RNA families with common functions have specifically conserved structural motifs, which are directly related to the functional roles of RNA in catalysis and regulation. Because the conserved structures depend on base-pairing, similar RNA structures may have little or no detectable sequence similarity, making the identification of conserved RNAs difficult. This is a particularly serious problem when studying regulatory structures in RNA. In many cases, such as that of cellular internal ribosome entry sites (2), although we can identify RNAs that have similar regulatory responses, it is difficult to tell whether the RNAs have common structural features using current methods. Available tools for identifying common structures based on RNA sequence suffer from one or more of the following problems: they do not consider pseudoknots, which are important

in many catalytic and regulatory structures; they do not consider near minimum free energy structures, which is important as many RNAs exist as an ensemble of structures of nearly equal energy; they require many examples of known structures in order to train a computational model; they require impractical amounts of computational time, precluding their use on long sequences or genomic scale; or they use a similarity function that cannot identify RNAs as having similar structure, even when they are from one of the well characterized known classes. The approach presented here has the potential to address all of these issues, allowing novel RNA structures that are shared between RNAs with little or no sequence similarity to be discovered. This provides a powerful tool to investigate and explain the pervasive transcription observed in eukaryotic genomes (1).

CHAPTER 1. INTRODUCTION

1.1 RNA structure: base pairing and structural elements

The traditional definition of RNA secondary structure is built on base interactions, *e.g.* base-pairing and base-pair stacking. There are two types of base-pairs: canonical and non-canonical base-pairs. Canonical base pairs include Watson-Crick base-pairs (A::U and G::C) and wobble base-pairs (mostly G::U, sometimes I::A, I::U, or I::C); the wobble base-pairs and mismatch pairs are referred to as non-canonical base pairs (3,4). Base-pair stacking refers to the situation in which two or more base-pairs stack on top of one another. Base-pairing and base-pair stacking produce stems, which are double-helical regions with at least two consecutive base-pairs. Unpaired bases form loops (Figure 1.1), which are single regions usually composed of at least three unpaired bases. Loops are typically divided into several types, depending on their position and topology of the unpaired regions: hairpin loops (also called stem loops), bulge loops, internal loops, multi-loops (also called junction loops) (Figure 1.1). Stems and loops are basic structural elements in RNA secondary structures. Pseudoknots are another relevant structural element. A pseudoknot is formed when bases in the loop of a stem interact with bases outside of this stem and form another stem (Figure 1.1). There are many variations of pseudoknots depending on the topology of the stems, the classical type being H-type pseudoknots (5) (Figure 1.2). Other types include kissing hairpins and hairpin-bulge

pseudoknots (Figure 1.2). Pseudoknots are found in many functional RNA molecules, including self-splicing introns, tmRNA, and RNase P RNA, etc., and they are considered to be important RNA structural elements, and are often associated with catalytic functions (6).

1.2 RNA's double life

Prior to the exciting discovery of catalytic RNA molecules four decades ago, people were reluctant to look beyond RNA's role as an information carrier in cellular systems. In the 1970s, Sidney Altman found, while studying tRNA biosynthesis, that the RNA component of ribonuclease P (RNase P) is essential for its enzymatic function, which is to cleave the mature tRNA part from the precursor sequence (7,8). Independent research by Thomas Cech revealed that the RNA component of the ribosome is essential for protein synthesis (9,10). Altman and Cech shared the 1989 Nobel Prize in Chemistry because of their discovery of catalytic RNA molecules. This was the beginning of the unveiling of RNA's double life, with later breakthroughs identifying additional catalytic and regulatory RNA molecules. In addition to rRNA, tRNA, and RNase, there are many other examples of functional RNAs, such as transfer-messenger RNA, self-splicing introns, riboswitches, attenuators, miRNA or siRNA in RNA interference, and CRISPR RNA. The emerging facts have shown that RNA is not only the intermediate molecule between DNA and protein but also a key player in many biochemical reactions (11).

1.2.1 RNA world hypothesis

The exciting breakthroughs in studies of functional RNAs have also given birth to the "RNA world" hypothesis, as an explanation of the origins of life on Earth (12). Under this hypothesis, in a very primitive system RNA plays the roles of both storing information and catalyzing biochemical reactions, with no DNA or protein required. For example, the disrupted function of miRNAs, or the differential expression of long non-coding RNAs, plays an important role in the initiation of human cancer (13). In addition, the RNA component in telomerase (14), is the major player in enzymatic functions. Research on functional RNAs have become relevant topics for studies in life science; the application of functional RNAs have become powerful tools in curing human diseases.

Ribosome/tRNA

The ribosome, tRNA, and mRNA, together form a protein synthesis factory. The ribosome is a sub-cellular component that can be found in all living cells. It consists of two subunits, large and small, each of them contains both RNAs and proteins. The RNA in the ribosome is called rRNA, which is the key component for ribosome function. Messenger RNA (mRNA), is an RNA chain transcribed from DNA. The protein coding sequence in mRNA is a sequence of nucleotide triplets called codons. Transfer RNA (tRNA), is an RNA molecule folded into an "L" shape in the three dimensional space. See Figure 1.3 for the tertiary and secondary structures of tRNA. The "upper stem" of the cloverleaf is called acceptor stem; the "lower stem" and the "lower leaf" of the cloverleaf are called anticodon arm and anticodon loop; the two remaining "stems" on the two sides are called D

arm and T arm and their two corresponding “leaves” are called D loop and TΨC loop (9,10). The anticodon loop contains a triplet of nucleotides called the anticodon. tRNA can be attached to amino acids; a tRNA molecule with a specific anticodon can only be attached to a specific amino acid. Translation is accomplished by the complementarity between anticodons and codons. In the ribosome, the small subunit binds to the mRNA, and the large subunit gathers the amino acids carried by tRNA and assembles the poly-peptide chain according to the order of triplet codons specified by mRNA. Ribosomes from different domains, for example, eukaryotes and prokaryotes, differ in size in each particle in the subunits; however, they share a conserved core with common folded RNA structures and carry the same function: protein biosynthesis (11).

RNase P

Ribonucleases (RNase) are enzymes that catalyze the cleavage and degradation of RNA molecules. Most RNases are proteins; however, RNase P is a ribozyme, which means it contains a catalytic RNA molecule. RNase P catalyzes the generation of mature tRNAs by cleavage of pre-tRNAs (12,13). RNase P is found in all three kingdoms of life (bacteria, eukaryotes, and archaea), and in protein-synthesizing organelles (mitochondria and chloroplasts). The RNase P holoenzyme consists of both RNA and protein, with RNA act-ing as the catalytic core and the protein providing support for the enzyme function. In RNase P, the structure of the catalytic RNA is conserved across three kingdoms of life (14). However, the protein differs in structural complexity: bacteria < archaea < eukary-otes, and some RNase P proteins in eukaryotes are responsible for catalytic function as

well (15). The interesting functional variation in the RNA and protein in RNase P, across species, agrees with what has been suggested in the "RNA world" hypothesis. RNA is the major player in basic cellular processes such as transcription and translation.

Group I & II introns/spliceosome

Group II self-splicing introns are a group of ribozymes found in all three kingdoms of life. They catalyze their own cleavage from host genes, such as tRNA, rRNA, and mRNA in chloroplasts and mitochondria. Group II introns contain 6 domains, labeled I to VI, with domain V being structurally conserved and functionally critical (15). The spliceosome is a complex cellular machine that catalyzes mRNA splicing in eukaryotes by removal of introns from the pre-mRNA sequence. The spliceosome catalyzes splicing through mechanism identical to the Group II intron. It is composed of ~60 to 150 different proteins and 5 small nuclear RNA (snRNA) molecules: U1, U2, U4, U5, and U6, with U2 and U6 in the active site (16). The spliceosome U2/U6 snRNA is highly similar to domain V in Group II introns, in both sequence and structure. The similarities between the spliceosome and Group II introns in function, sequence, and structure, has led to the hypothesis that the spliceosome has evolved from Group II self-splicing introns (16). Another group of self-splicing introns, called the Group I introns, catalyze their own excision from tRNA, rRNA, and mRNA precursors in a variety of organisms including bacteria and eukaryotes (16). The active site of Group I intron is a conserved core composed of two helical domains made from paired RNA regions (P): P4-P5-P6 and P3-P7-P9 (17,18).

tmRNA

Transfer-messenger RNA, tmRNA, is a bacterial RNA molecule with a dual function as both a tRNA and an mRNA. tmRNA forms the tmRNP complex with Small Protein B (SmpB) and performs trans-translation. When an mRNA lacking stop codons is being translated, the ribosome may get stalled and produce a truncated polypeptide. tmRNA releases the stalled ribosome, adds a proteolysis-inducing protein tag to the end of the truncated protein, and facilitates the degradation of the mRNA involved in the stalled ribosome (19). The functional core of the tmRNA structure is composed of a tRNA-like domain (TLD) and a mRNA-like domain (MLD), connected by a pseudoknot-rich domain (PKD). The MLD contains a short open reading frame (ORF) encoding the degradation-inducing protein tag (a short polypeptide chain that guides degradation by housekeeping proteases), with resume and stop codons surrounding the ORF. The tmRNA resumes and then finishes translation of the nonstop mRNA by moving the ribosome onto the resume codon in the tmRNA and continuing translation of the proteolysis-inducing protein tag encoded by the ORF (20). The structure of tmRNA has been obtained by comparative sequence analysis of aligned tmRNA sequences from multiple organism based on covariance in base pairing (21,22). Although there is no available full-length tmRNA structure at the atomic level, crystal or cryo-EM structures of TLD have been solved recently (23,24).

Attenuator/riboswitch

Attenuators are base-paired RNA elements found in the 5' untranslated regions (UTRs) of bacterial genes that sense environmental change and regulate gene expression according to environmental conditions. Environmental conditions include temperature and the concentrations of metabolites and macromolecules (23). The major types of attenuators include riboswitches, T-boxes, peptide leaders, ribosomal protein leaders, and binding sites of terminators, and anti-terminator proteins (24). Riboswitches are among the most well-known attenuators. Riboswitches are natural RNA aptamers (an unpaired region of RNA sequence with high affinity to a specific metabolite) that are found in both Gram-positive and Gram-negative bacteria (25). Riboswitches are embedded in the 5' untranslated region (UTR) of genes involved in the production of specific metabolites, and control the transcription or translation of metabolites by switching their structure upon the binding of a regulatory ligand. The metabolic pathways that are affected include biosynthesis of vitamins, metabolism of amino acids, and metabolism of nucleobases (26). A typical riboswitch contains two parts: an RNA aptamer and an expression platform (a folded region of the RNA functioning as a transcription terminator). The aptamer is highly conserved across species for the same class of riboswitches. The aptamer selectively binds to the metabolites and changes the structure of the expression platform upon binding. This forms a terminator stem that terminates transcription by binding a terminator protein, or sequesters the ribosome-binding site and prevents initiation of translation (17). Riboswitches are ancient mechanisms for regulation of gene expression (18).

miRNA/siRNA

Micro-RNAs (miRNA), or small interfering RNAs (siRNA), are short (miRNA, ~21-25 nucleotides, siRNA, ~20-24 nucleotides) regulatory RNA molecules that are responsible for sequence-specific gene silencing, which is also known as RNA interference (RNAi) (27). MiRNAs are derived from precursor RNA molecules, which are transcribed from genomic regions encoding the genes to be silenced; these precursor molecules are stem-loop structures, and part of their sequences contain the sequence of the miRNA. By contrast, siRNAs are derived either from infecting viruses or artificial synthesis, (exogenous), or derived from aberrant transcripts (endogenous)(28-30). The processing of both miRNAs and siRNAs include multiple steps: processing into small RNA duplexes by an RNaseIII enzyme called Dicer, unwinding into single stranded RNAs (ssRNAs), loading of one strand into the RNA-induced silencing complex (RISC), guiding of the RISC to the target transcript (complementary to the ssRNA), and degradation of the target transcript by a family of endonucleases in the RISC called Argonaute (31). Because of its use for knocking down expression of target genes, RNAi induced by siRNAs and miRNAs have multiple uses including high-throughput studies of gene regulation, cure of viral infections, and hopefully, development of other disease therapeutics. For example, RNAi can be used to silence genes that are differentially expressed in tumor cells as cancer therapeutics.

1.2.2 RNA structure and function conservation

Functional RNA molecules are folded into complex structures and involved in multiple important cellular processes, such as transcriptional and translational regulation (19). The function of these RNA molecules depends on the presence of conserved motifs. As only a small number of functional RNA species have been catalogued so far, the majority of functional RNA motifs are yet to be identified. For example, in the human ENCODE (Encyclopedia of DNA Elements) project, the function of most of the small RNAs are yet to be confirmed (25).

In DNA and protein, traditional approaches used to detect conserved functional motifs are based on sequence similarity; however, the low sequence similarity in functional RNAs makes it difficult to identify functional motifs in RNA based on sequence similarity alone (21). Despite the possible lack of sequence conservation, RNAs with similar functions typically have conserved secondary or tertiary structures (22), which offers an alternative approach for identifying functional elements in RNA – conserved structural motif identification.

Unlike DNA and protein, in which conserved motifs are encoded on the primary sequence level, regulatory and catalytic motifs in RNA are base-paired structures. The topological arrangement of these structures, for instance the nesting of stems, multi-loops, and pseudoknots, is critical to the structure and function of the molecule. RNAs with similar functions, for example those in RNase P, the ribosome, or self-splicing introns, typically have strongly conserved topologies (15,26-28). The importance of iden-

tifying RNAs with similar topologies is therefore comparable to the importance of sequence alignments in identifying conserved protein and DNA structures. One of the notable aspects of RNA structure is the importance of pseudoknots. For example, in the self-cleaving ribozyme in Hepatitis Delta Virus (HDV), a double-pseudoknotted structure forms the catalytic core which is critical for viral infection (6) (Figure 1.4); in Group I Self-Splicing Introns, pseudoknots form the catalytic core of the splicing reaction (29,30). Therefore, the identification of conserved topologies that include pseudoknots may be critical to identifying biologically important structures.

1.3 Current approaches to study of RNA structures

1.3.1 Experimental

1.3.1.1 Individual RNA molecules: X-ray, NMR, and chemical/enzymatic probing

Common experimental approaches for study of RNA structures include biophysical tools such as X-ray crystallography, NMR spectroscopy, and probing using enzymes or chemicals. RNA structures can be accurately determined by crystallographic or NMR approaches, but these approaches remain very difficult – only a few hundred large RNA structures have been determined, and most of these belong to one of only a few classes. RNA structure probing, also known as RNA structure footprinting, in which RNA structures are cleaved at specific positions, *e.g.*, at paired or unpaired regions, can be used to determine which bases in a structure are paired, but not the bases to which they pair. Chemical probing approaches include DMS probing, CMCT probing, Kethoxal probing,

and SHAPE. Enzymatic probing of RNA structures uses nucleases such as RNase V1, S1 nuclease, RNase T1, RNase T2, etc. Chemical or enzymatic probing is followed by primer extension and gel electrophoresis to determine the location of cutting, and with the known base specificity of the reagent, the information of pairing of a specific base can be indirectly inferred. The procedures of chemical or enzymatic probing will be introduced below.

DMS/CMCT/Kethoxal

Dimethyl sulfate (DMS), is a chemical reagent that modifies A bases and C bases by methylation of their base-pairing faces. Base-pairing protects the bases from methylation. Using DMS methylation followed by primer extension and gel electrophoresis, the unpaired As or Cs on an RNA sequence can be detected, as the primer extension stops at the DMS-methylated base (31). Similar to DMS, chemical reagents that modify other RNA bases also have been applied in RNA structure probing. 1-cylcohexyl-(2-morpholinoethyl) carbodiimide Metho-p-toluenesulfonate (CMCT) modifies unpaired Us by alkylation (32,33). Kethoxal modifies unpaired Gs by alkylation (34). One drawback of these chemical probing methods is that they are only sensitive to one or two bases; therefore, sometimes these three chemicals are combined for a complete analysis of RNA local structures (35). The data obtained by DMS/CMCT/Kethoxal probing can be used as constraints in RNA structure prediction for higher accuracy (36). However, probing using the different chemicals needs to done separately, and combining data

from multiple experiments creates noise. Moreover, differences in experimental conditions affect the RNA structure.

SHAPE

Selective 2'-Hydroxyl Acylation and Primer Extension (SHAPE), is an approach that uses the selective chemical reactivity of the RNA ribose hydroxyl group for identification of base-paired and non-base-paired regions in RNA. The chemical reactivity of the RNA ribose 2'-hydroxyl group is sensitive to the presence of an adjacent 3'-phosphdiester. For a paired base, the 2'-hydroxyl group is strongly constrained by its neighboring 3'-phosphodiester anion and thus unreactive. Therefore, unpaired bases are more likely to be chemically reactive. In a SHAPE experiment, RNA molecules are probed with a 2'-hydroxyl reactive chemical reagent, such as N-methylisatoic anhydride (NMIA); the chemically reactive bases are then visualized and quantified by primer extension and gel electrophoresis, which identifies the pairing status of specific bases. Using SHAPE, quantitative nucleotide-resolution local RNA structures or maps of paired bases can be obtained (37,38). Similar to the application of DMS/CMCT/Kethoxal probing data in RNA structure prediction (36), the base pairing information obtained from SHAPE experiments can be used as pseudo-free energies and incorporated into nearest neighbor model based energy minimization to improve structure prediction accuracy (39).

Enzymatic probing

Similar to chemical probing, enzymatic probing is performed by digesting RNA with nucleases with different specificities, such as for unpaired regions (S1 nuclease, RNase T1, RNase T2, RNase A, RNase U2) or paired regions (RNase V1) (40).

In general, a major challenge for all the probing approaches is that the probed structure is an average of the repertoire of structures, which may not be the conformation of a real structure, since multiple conformations could exist simultaneously. Another issue with the probing approaches is that they might not be accurate as they are usually done on extracted RNA, which differs from the RNA in living cells. The *in vivo* environment is more complex than the *in vitro* environment; change in solution conditions, and the binding of metal ions and proteins could change the RNA structure drastically. Furthermore, the success rate of the probing experiment is highly sensitive to the reaction time of the reagent, as over digestion creates fragmented RNA sequences with low specificity in the following sequencing, and insufficient time can cause no digestion at all. In addition, the experimental procedures, such as cell lysis and RNA extraction, further decrease the stability of RNA structures and create fragments of RNA sequences that adds to the noise (41).

1.3.1.2 Multiple RNA molecules: transcriptome chemical/enzymatic probing and high-throughput sequencing

While classical chemical/enzymatic probing focused on only one RNA at a time, simultaneous or genome-wide probing of multiple RNAs has emerged in recent years. In these experiments, chemical or enzymatic probing is performed on the transcriptome, followed by quantification using high-throughput sequencing. These experiments provide not only a new approach to RNA structure determination, but also provide high-throughput data that improves computational methods for RNA structure prediction.

PARS/FragSeq

Examples of high-throughput enzymatic probing include parallel analysis of RNA structure (PARS) in yeast (42) and fragmentation sequencing (FragSeq) in mouse (43). In these experiments, transcripts were digested with structure-specific enzymes; single-stranded or double-stranded positions were identified by subsequent sequencing, as the sequencing could reveal the positions of specific enzyme cutting.

DMS-seq/SHAPE-seq

Examples of high-throughput chemical probing include DMS-seq (44) and SHAPE-seq (45). In these experiments, transcripts are treated with chemicals probing unprotected bases, with or without base specificity. The limitations of probing methods still exist, such as the inaccuracy of RNA structures due to the averaging of conformations, and the noise created by the inability to control in the extent of reaction with the probing rea-

gent. Moreover, pursuit of high throughput data may sacrifice the accuracy of RNA structures, as it is difficult to customize the experimental conditions for each RNA.

1.3.2 Computational

Other than experimental strategies, RNA structure analysis heavily relies on two computational approaches – structure prediction and covariance analysis.

1.3.2.1 RNA conformational free-energy parameters for structure prediction

The prediction of RNA secondary structures is based on prediction of the minimal free energy (MFE) structure. The energy minimization is based on a nearest-neighbor model, which computes the free energy of bases using information about adjacent paired, or stacked, bases (46). The nearest neighbor model is an approximation of the RNA folding stability of RNA secondary structures. In the nearest neighbor model, the stability of an RNA sequence is calculated by combining the experimentally determined thermodynamic parameters of structural elements such as helices, loops, and pseudoknots, following a set of thermodynamic rules. In the nearest neighbor model, an RNA duplex is decomposed into a series of base-pair doublets, and its free energy is calculated as the sum of the free energies of those doublets, plus additional terms such as duplex (helix) initiation/termination and penalties for some unstable structures, such as loops, stems ending in A::U or U::A base-pairs, or hairpin loops formed by only C bases (47-49). The original thermodynamic parameters, which were obtained from melting experiments dating back to the 1970s (50-52), have been extended as more experimental data be-

came available in the 1990s (48,53). In the 2000s, chemical probing constraints were incorporated into the thermodynamic parameters, which increased the accuracy of RNA structure prediction (36). A database collecting the existing thermodynamic parameters was established in 2010 (54).

1.3.2.2 RNA structure prediction programs

In RNA folding, the primary assumption is that the most favorable structure is the one with the minimum free energy (MFE). However, RNA structures are dynamic and have interconverting states in folding. In addition, RNA structures are affected by environmental conditions such as salt concentration, temperature, and binding of proteins. An RNA structure is an ensemble of structures with near-MFE energies and different conformations (55). Given the number of possible conformations within an energy range, obtaining all the possible near-MFE structures is computationally expensive. Dynamic programming (DP) algorithms (56-61) using experimentally-determined nearest-neighbor thermodynamic parameters (54) (see chapter 1.3.1.2 for more details) are widely used to calculate minimum and near-minimum free energies of RNA folding. To limit the number of predicted conformations, some programs incorporate the McCaskill partition function algorithm to sample RNA conformations from the Boltzmann distribution by probability. UNAFold (3,62,63) is one of the most popular dynamic programming-based RNA structure prediction approaches. UNAFold computes the thermodynamically optimal structure using dynamic programming; this work can be extended to

identify suboptimal structures within a range of free energy of the optimal folding. RNAstructure (64-66) and ViennaRNA (67-69) are also based on dynamic programming, and have incorporated partition function calculation for computation of base-pairing probabilities. Another example of a partition function-based program is Sfold (70,71), which calculates a centroid structure based on the base-pairing probabilities of a Boltzmann ensemble. Programs such as CONTRAfold (72,73), CentroidFold (74) and IPknot (75), calculate base-pairing probabilities using conditional log-linear models (CLLM), a grammar-based method similar to stochastic context-free grammars (SCFG).

In general, dynamic programming based programs cannot predict pseudoknots due to computational complexity. The prediction of pseudoknots requires calculation of non-nested base-pairs, which is a NP-hard problem (76). To solve this problem, some programs have limited the range of parameters in their dynamic programming algorithms, or applied heuristics to predict only certain types of pseudoknots. These programs include RNAPKplex (77) in the ViennaRNA package, ProbKnot in the RNAstructure package (78), DotKnot (79,80), and pKiss (81,82).

In general, due to memory and time limitations, dynamic programming based programs do not predict pseudoknots; however, some program suites have extended secondary-structure prediction to include pseudoknots by incorporating various heuristics into their algorithms. ViennaRNA includes the program RNAPKplex (77), which decomposes a secondary structure into two parts and separately calculates the minimum free energy of each part. The two parts include a pseudoknot-free structure that includes accessible (unpaired) bases, and an additional stem formed within the accessible region to form a

pseudoknot with a stem in the pseudoknot-free structure. The calculation of pseudoknot energy is recursive and with complexity $O(n^6)$, and n is the length of sequence of base-pair number; when the length of the accessible region limited to w, the computational time for RNAPKplex is $O(n^3 + n^2w^4)$. RNAPKplex decomposes a secondary structure into two parts, a pseudoknot-free structure with unpaired bases, and an additional stem formed using the unpaired bases to form a pseudoknot with a stem in the pseudoknot-free structure, and computes the free energies separately. ProbKnot calculates a maximum expected accuracy structure based on the partition function, and that structure may or may not contain pseudoknots. DotKnot predicts pseudoknots by assembly of high probability base-pairs and evaluation of those base-pairs using pre-determined pseudoknot energy parameters. pKiss predicts pseudoknots using heuristics. DotKnot has a high precision, around 80%, in predicting pseudoknots in short RNA sequences, while ProbKnot has a high precision, ranging between 60% and 80%, and a fluctuating recall, ranging between 50% and 90%, in longer RNA sequences. Due to the limitations in parameters or computations, these programs predict only certain types of pseudoknots, and their precision or recall still has some space for improvement, such as incorporation of pseudoknot energy parameters in further experiments.

1.3.2.3 Phylogenetic approaches

Another computational approach to determining RNA structure is phylogenetic analysis, which is based on the assumption that the sequence of functional RNAs remains un-

changed or changes simultaneously when compared to their surrounding regions. In phylogenetic approaches, homologous sequences from a diverse set of organisms are aligned, and bases in each sequence occupying the same column in the alignment are identified. There are two types of phylogenetic approaches: phylogenetic footprinting (83) and covariance analysis (84,85). Phylogenetic footprinting identifies conserved regions by comparing a candidate sequence to orthologous sequences in different species. The conserved regions are the identical regions in the global alignment across orthologous sequences, which are considered to be paired in the folded RNA sequence. It is convenient to identify conserved regions using phylogenetic footprinting, however, the conserved regions sometimes could be short (5-10 bases), when compared to the entire region being scanned (1000 bases), and might be covered by the non-functional regions. Covariance analysis works in a similar way, but the comparison is between species that are less related, and thus have less sequence conservation. Covariance is a phenomenon in which changes in the sequence at two separate positions coincide to maintain the base-paired structure. A certain number of base-pairs are consistently found in multiple sequence alignments, and change in one base causes corresponding changes in another so as to maintain base-pairing (86). Covariance approaches require the alignment of a set of sequences; these sequences usually come from the same gene of interest in related species. Base-paired regions can be identified as sequence positions that show complementary base changes, for instance, an A$\rightarrow$C base change is associated with a U$\rightarrow$G base change at the base-paired position. This is usually detected by calculating the mu-

tual information between positions in the aligned sequences (87). Covariance analysis is a major approach to use to validate the existence of predicted structures (84,85). An example of covariation is shown in Figure 1.5. Despite its advantages, covariance sequence analysis requires that homologous sequences be available for different species, and that the RNA sequences being examined to be highly conserved to discriminate the functional regions from the non-functional regions, which is not always the case. Furthermore, it can be computationally expensive if the region being searched is on the genomic level.

1.4 Visualization of RNA structures

Many approaches are available for visualization of RNA secondary structures. Simple approaches focus on annotation of RNA base-pairing, and some of them include the free energy information of RNA structures. The most widely used visualization methods include stem-loop diagrams, circle plots, dome plots, energy dot plots, the Vienna format notation from ViennaRNA package, the Connect format notation from the UNAFold program suite, and the BPSEQ format notation. The details about these representations will be presented below.

The first four simple representations mentioned above offer straight-forward visualization of RNA secondary structures. In stem-loop diagrams, the RNA sequence is plotted as a curved line with unpaired regions as loops and stems as ladders. The circle plot is a base-pairing annotation proposed by Nussinov in 1978 (88), where the RNA sequence is

represented as a circle with the position of each base indicated by dots. Two dots are connected by a line if the two corresponding bases are paired. The dome plot shows the sequence as a line with positions of the paired bases connected by arcs (also called domes). Therefore, the plot of a stem with multiple base-pairs has multiple domes. For simplification, multiple domes in the same stem can be represented by a single arc. The energy dot plot was proposed by Zuker for representing RNA secondary structures, and their predicted free energies, predicted by mfold (3,63). In the energy dot plot, the sequence is plotted against itself, with the position of each base-pair shown as a colored dot. One plot usually contains several alternative structures including the minimum free energy (MFE) structure. For better visualization, each alternative structure is labeled with its free energy and colored differently. See Figure 1.6 for an example of these simple visualizations of RNA structures.

The remaining three approaches mentioned in the beginning of this section, which are the Vienna format notation, the Connect format notation, and the BPSEQ notation, provide annotation of the defaults of base-pairing in RNA structure. The Vienna format notation, proposed in the ViennaRNA package (89), is also known as the dot-bracket representation. In the Vienna format, each base in an RNA secondary structure is represented as either a dot (unpaired) or half of a bracket (paired). In the Vienna format, one or multiple alternative structures for one RNA sequence are shown, and the difference between alternative structures can be shown by the differences between dots and brackets. The Connect format notation, an RNA structure representation proposed in mfold (3,63), also describes one or more structures corresponding to one sequence. It starts

with the name of the sequence and the free energy of one alternative structure (if available), followed by the base-pairing information for each base: the base content, its preceding base, its following base, and the base it is paired to. The BPSEQ formation notation is a succinct variation of the Connect format. See Figure 1.10 for an example of an RNA structure in these three formats.

1.5 Comparison of RNA structures

Graphical representation of RNA structures is intended for efficient representation and comparison of structures. RNA structures have been commonly represented as tree graphs, but only secondary structures (not pseudoknots) can be included (58,90-93). Another graphical representation approach has been implemented in the RNAshapes package by the Giegerich group (94-96), which represents RNA structures as abstract shapes. Shape abstraction retains nesting and adjacency in the structure, but removes information such as helix length, aiming for efficient computation. This approach is described in detail below. These approaches, however, do not include pseudoknots in either the representation or the analysis. The Schlick group proposed the RNA-As-Graphs (RAG) method, which represents RNA structures either as tree graphs (without pseudoknots) or dual graphs (with pseudoknots) (97-100).

1.5.1 RNAshapes

The RNAshapes package represents RNA structures within a certain folding space as abstract shapes, in which multiple base-pairs or unpaired bases are represented by single symbols, aiming for efficient RNA structure comparisons (94-96). In RNAshapes, unpaired regions are displayed as underscores, and stacking regions are displayed as pairs of brackets. According to the level of abstraction, several abstraction variations are available. In higher levels of abstraction some information describing nesting and adjacency is removed. Figure 1.7 shows an example with the lowest level of abstraction. The RNAshapes approach is limited to pseudoknot-free structures.

1.5.2 RNA-As-Graphs

The RNA-As-Graphs (RAG) method represents RNA structures as tree graphs without pseudoknots (Figure 1.8) or dual graphs with pseudoknots included (Figure 1.9) (97-100). An RAG is quantified by numerical descriptors, such as the eigenvalue spectrum of the Laplacian matrix, as a measurement of graph compactness and connectivity, and topological numbers as measurement of graph isomorphism (97,100). These numerical descriptors provide limited ability to identify similar structures. They have never been shown to be able to group RNAs into structural/functional classes, or to identify subgraphs nested within larger graphs (> 10 vertices), as a typical graph of RNA structure may have up to 20 vertices.

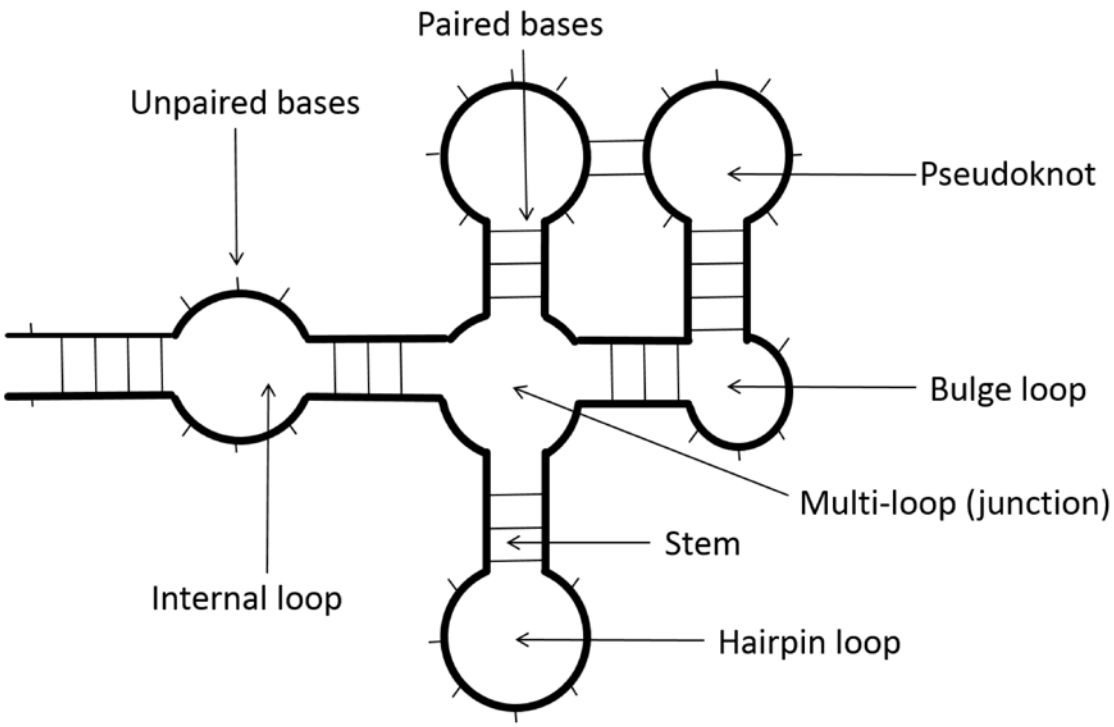
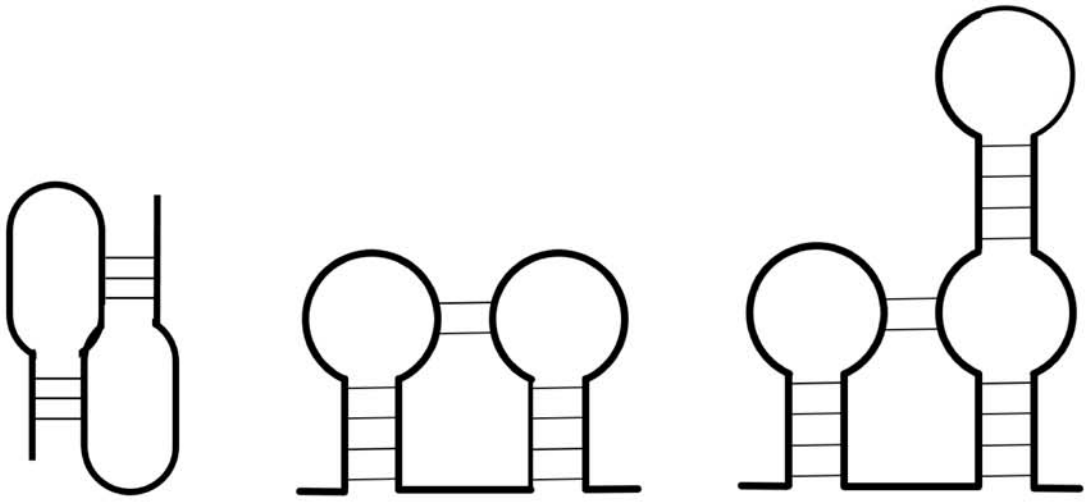


Figure 1.1 RNA structure with stems, loops, and pseudoknots. A stem is a double-helical region usually composed of at least two consecutive base-pairs. Loop is an unpaired region usually composed of at least three unpaired bases, as loops with less than three bases are unstable and usually cannot form. A pseudoknot is formed when bases in the loop of a stem interact with bases outside of this stem, and form another stem. A hairpin loop in RNA is defined as a loop enclosed by a stem. An internal loop in RNA is defined as a loop enclosed by two adjacent stems. A typical internal loop comprises two unpaired areas, one on each strand; however, a special case exists when the loop only contains an unpaired area on only one strand – a bulge loop. A multi-loop in RNA is defined as a stem enclosing two or more stems.



H-type pseudoknot Kissing pseudoknot Hairpin-bulge pseudoknot

Figure 1.2 Common types of pseudoknots. The bold lines indicate the sequences, and the thin lines indicate the base-pairing at specific locations on the sequence.

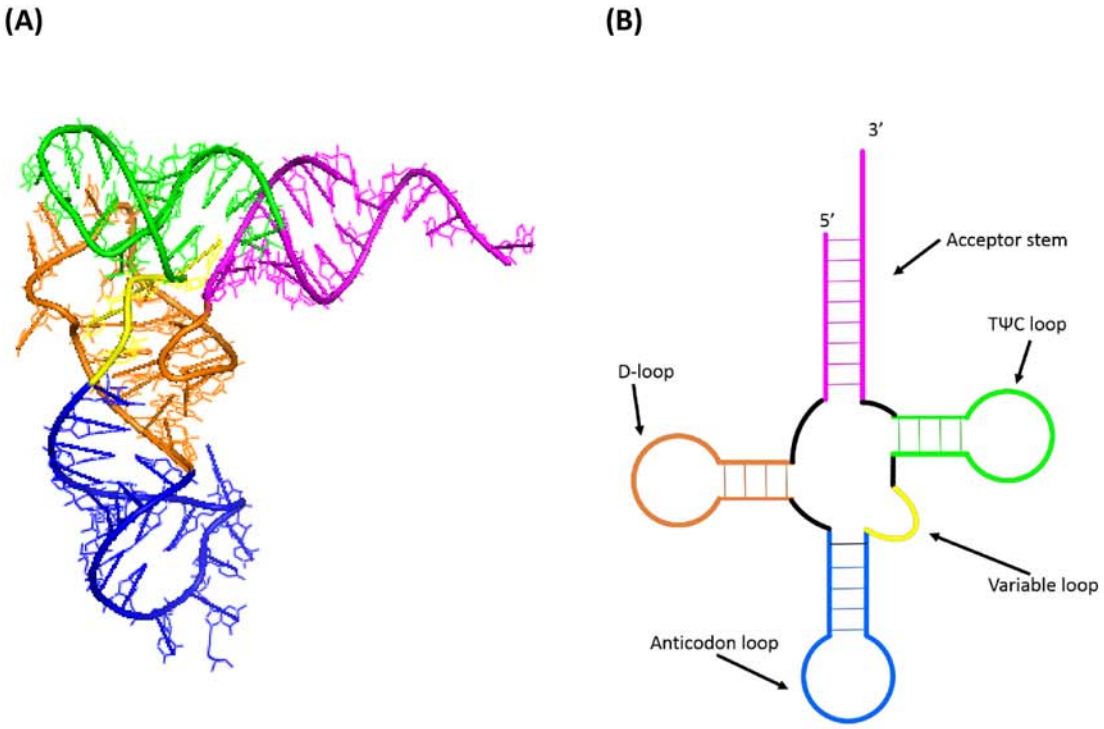


Figure 1.3 tRNA structure. (A) Tertiary structure; (B) Secondary structure. The stems and loops in (B) correspond to structural motifs in (A), as indicated by the color.

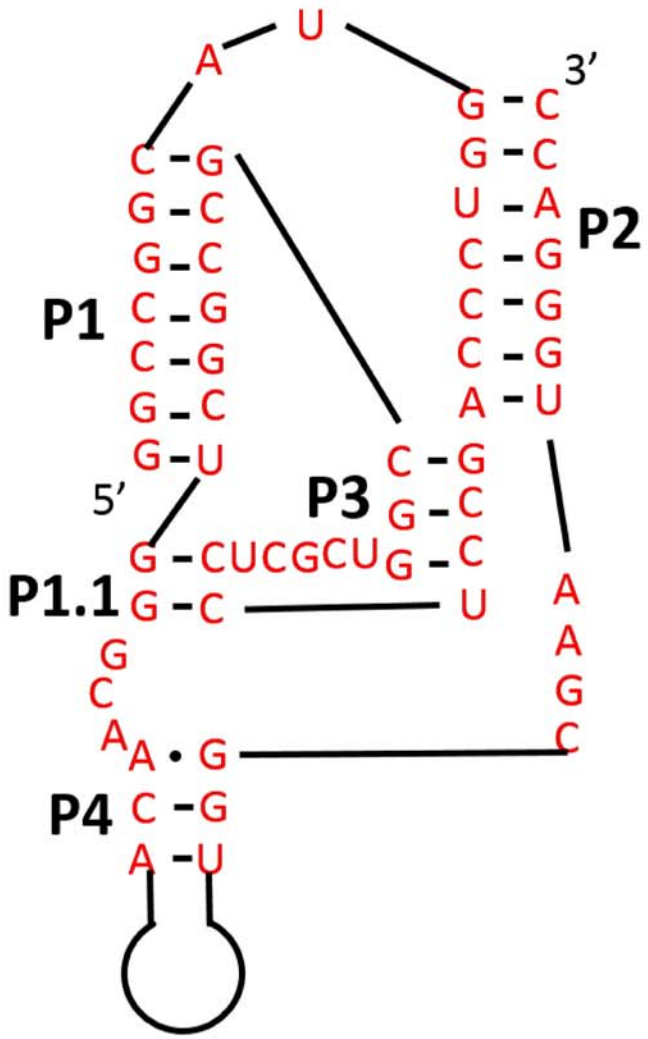


Figure 1.4 Structure of the self-cleaving ribozyme in Hepatitis Delta Virus (HDV). The catalytic core of this ribozyme is composed of two pseudoknots: P1-P2, and P1.1-P3.

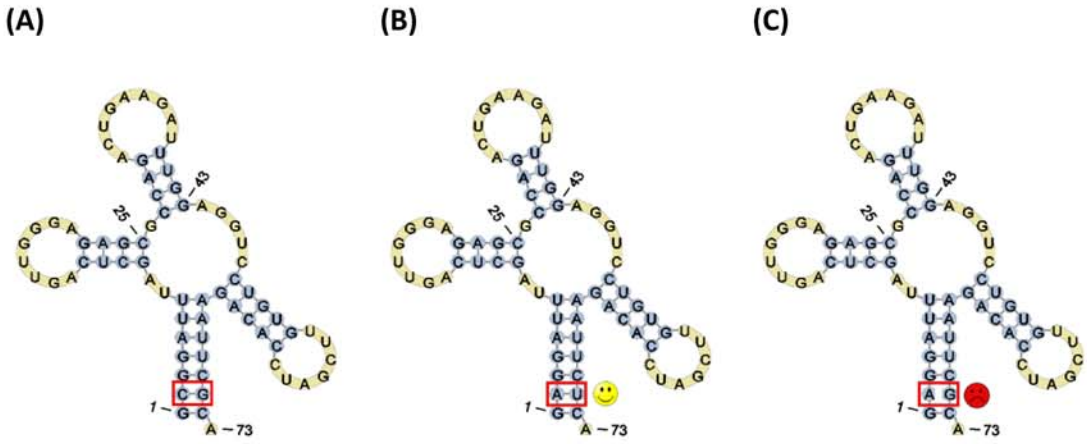


Figure 1.5 An example of covariation in RNA structure. (A) The original structure; (B) Co-variation; simultaneous change of C and G into A and U, as indicated by the smiley face, which ensures that base-pairing is maintained, and the RNA structure is conserved; (C) Mutation in one base, C changed to A, but G stays the same, as indicated by the sad face, which makes base-pairing impossible and destabilizes the structure. The visualization was created with PseudoViewer3 (101).

Figure 1.6 Simple representations of RNA structure.The diagrams show two predicted structures: the predicted minimum free-energy (MFE) (-27.10 kcal/mol) structure, and a near-MFE (-23.78 kcal/mol) structure predicted for Staphylococcus aureus tRNA-Isoleucine (102) using UNAFold (63). (A) The stem-loop diagrams of the two alternatives structures, created by the RNAstructure online server (64); (B) The circle plots of the MFE and near-MFE structures, created by Matlab; (C) The dome plot of the MFE and near-MFE structures, created by Matlab; (D) The energy dot plot of the MFE (red) and near-MFE structures (black), created by UNAFold (63).

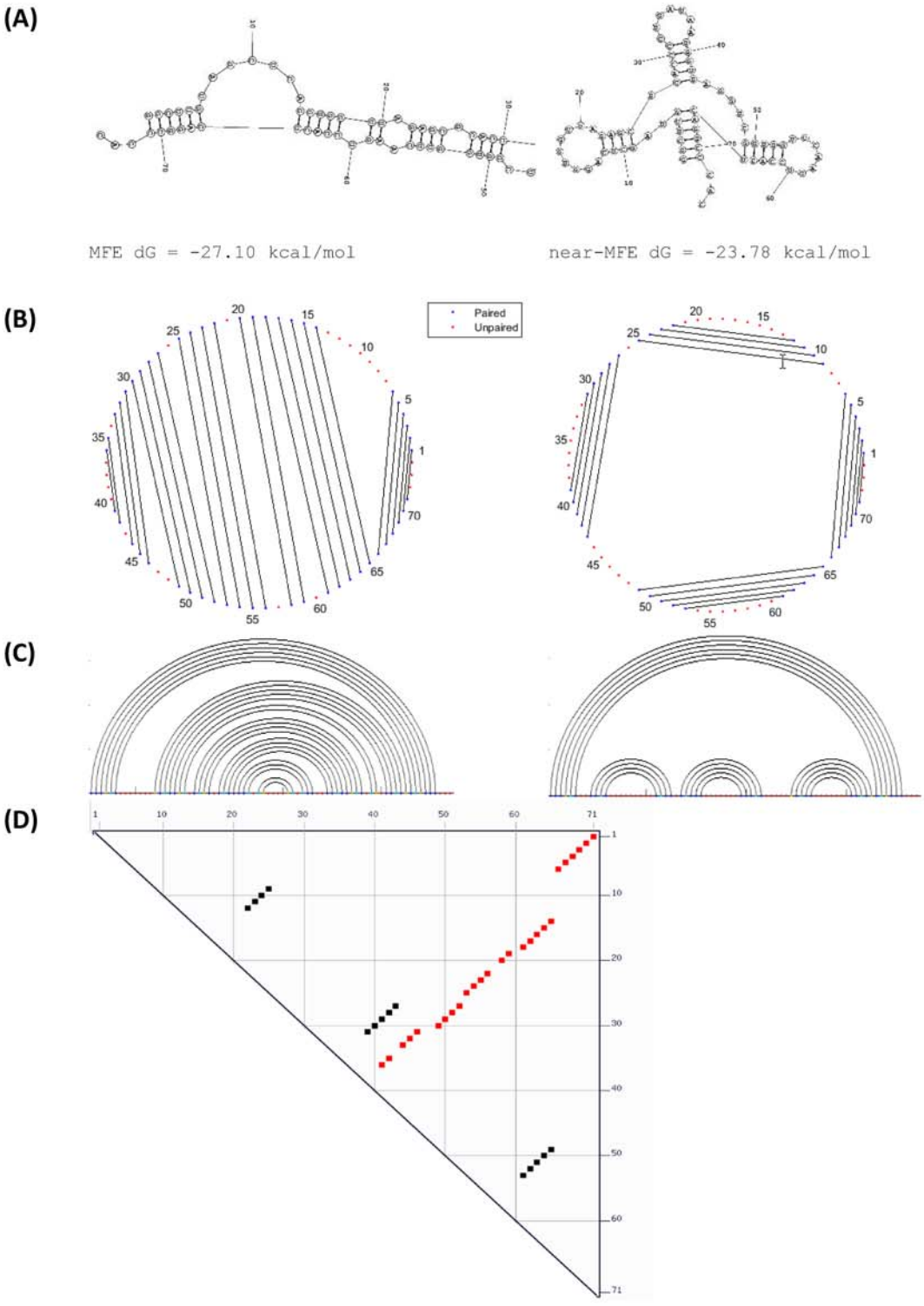

(A)
MFE dG = -27.10 kcal/mol
near-MFE dG = -23.78 kcal/mol
(B)
Paired
Unpaired
(C)
(D)

Figure 1.6

(A)

```
>tRNA.1QU2.fa
GGCUUGUAGCUCAGGUGGUUAGAGCGCACCCCUGAUAAGGGUGAGGUCGGUGGUUCAAGUCCACUCAGGCCCAC
((((((.......(((((((.((((.(((((((.((....)).))).)))))))).)).)))))))))...  (-27.10)
```

(B)

```
[_[_[_[_[_]_]_]_]_]_
```

Figure 1.7 RNAshapes representation. This figure shows the secondary structure represented in (A) ViennaRNA format and (B) RNAshapes format (least abstraction level), respectively. In RNAshapes format, the unpaired regions are represented as underscores and stacking regions as square brackets.

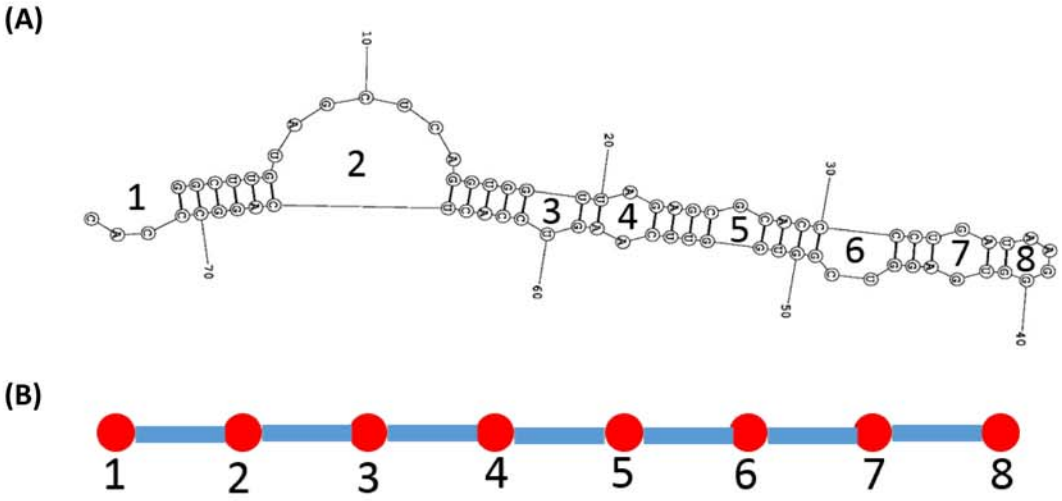
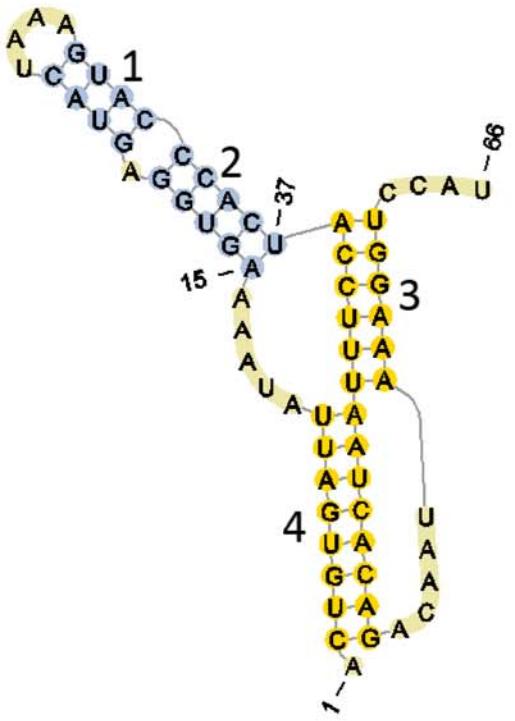


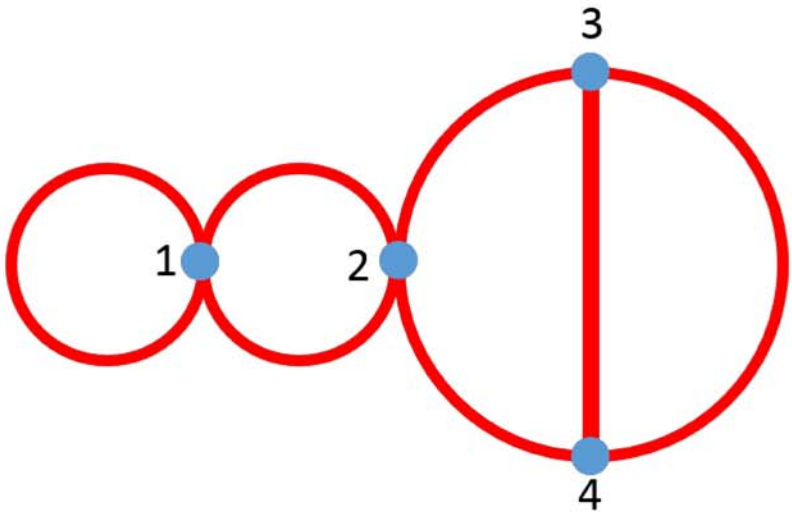
Figure 1.8 RNA-As-Graphs (RAG) tree graph representation. The figure shows the secondary structure represented as (A) a stem-loop diagram and (B) an RAG tree graph. In the RAG tree graph, loops (unpaired regions with more than 1 unpaired base) are represented as vertices (1-8) and stems (stacking regions) as edges. The corresponding loops in (A) and vertices in (B) are labeled with the same numbers. The stem-loop visualization is made by RNAstructure (64).

Figure 1.9 RAG dual graph representation. A 3'-terminal pseudoknot in strawberry chlorotic fleck associated virus is determined by comparative sequence analysis (103). This figure shows the secondary structure as a stem-loop diagram (A), created by PseudoViewer3.0 (101), RAG dual graph (B), sequence (C), and ViennaRNA format (D), respectively. In the RAG dual graph, stems (stacking regions with more than 1 base-pair) are represented as vertices and loops (unpaired regions) as edges. The corresponding stems in (A) and vertices in (B) are labeled with the same numbers.

(A)



(B)



(C)

ACUGUGAUUAUAAAAGUGGAGUACUAAAGUACCCACUACCUUUAAUCACAGACAAUAAAGGUCCAU

(D)

.(((((((.....(((((.((((....)))))))))[[[[[[)))))))).....]]]]]]....

Figure 1.9

Figure 1.10 Storage formats of RNA structures. (A) ViennaRNA format: each unpaired base is represented as a dot, and paired bases as a pair of matching brackets. (B) Connect format: the header line contains the sequence length, the free energy of a folding, and the sequence id. The following lines show the position of the base, base identity, position of the previous base, position of the following base, position of base that this base is paired to (0 if the base is unpaired), and a redundant index of this base which is the same as the first column. (C) BPSEQ format: a simplified version of Connect format, everything is the same except that it only contains the columns of the position of the base, base identity, and the position of base that this base is paired to.

(A)

```
>tRNA.1QU2
GGCUUGUAGCUCAGGUGGUUAGAGCGCACCCCUGAUAAGGGUGAGGUCGGUGGUUCAAGUCCACUCAGGCCCAC
((((((.......(((((((.((((.(((((((.((....)).))))..)))))))).)).)))))))))...
```

(B)

```
74      dG = -27.30 tRNA.1QU2
1       G       0       2       71      1
2       G       1       3       70      2
3       C       2       4       69      3
4       U       3       5       68      4
5       U       4       6       67      5
6       G       5       7       66      6
7       U       6       8       0       7
8       A       7       9       0       8
9       G       8       10      0       9
10      C       9       11      0       10
                    ...
68      G       67      69      4       68
69      G       68      70      3       69
70      C       69      71      2       70
71      C       70      72      1       71
72      C       71      73      0       72
73      A       72      74      0       73
74      C       73      0       0       74
```

(C)

```
74      dG = -27.30 tRNA.1QU2
1       G       71
2       G       70
3       C       69
4       U       68
5       U       67
6       G       66
7       U       0
8       A       0
9       G       0
10      C       0
            ...
68      G       4
69      G       3
70      C       2
71      C       1
72      C       0
73      A       0
74      C       0
```

Figure 1.10

CHAPTER 2. XIOS RNA GRAPH MATCHING

2.1 RNA XIOS graphs and XIOS format

As mentioned in the previous chapter, currently existing methods for RNA structure matching, either lack the ability to include pseudoknots, or are incapable of grouping RNAs into structural/functional classes. In order to deal with these limitations of current methods, we have developed a graph theoretic method that aims for efficient pattern matching of RNA structures.

2.1.1 XIOS graphical representation

One way to compare the topological relationships in the RNA structures is to identify conserved structural motifs within a group of RNA molecules using a graphical representation. We have developed the RNA XIOS graph theoretic approach (104) in which stems are represented as vertices, and stem-stem topological relationships as edges. Four types of edges are possible: **eXclusive** (two stems cannot form simultaneously because they share the same range of the sequence), **Included** (one stem is nested in another), **Overlapping** (two stems form a pseudoknot, *i.e.*, bases in the loop region of one stem interact with bases outside the stem, which generates another stem), and **Serial** (two neighboring stems form simultaneously and have no overlap) (Figure 2.1).

The XIOS graph approach has several advantages. (1) It can incorporate multiple structures into one graph, allowing the representation of a near-MFE ensemble in a single graph; studying the set of structures in an ensemble not only allows the detection of switch-like structures that change upon the binding of ligands (105), but also helps to identify pseudoknots (106). (2) XIOS graphs are designed to represent pseudoknots, which are one of the four possible relationships defined between stems. (3) All biologically possible XIOS graphs, and hence all biologically possible topologies, can be enumerated, and this enumerated set used to rapidly identify conserved structures (see Table 2.1).

2.1.2 XIOS format

The XIOS graphs are described in XML (Extensible Markup Language) format. Figure 2.2 shows an example of a XIOS file of RNase P RNA from *A.fulgidus*.

- An XIOS file contains four blocks: information (metadata), stem list, edge list, and adjacency matrix.
- The information block contains the RNA graph id, its functional category, and sequence information.
- The stem list shows each stem with its starting and ending positions in the sequence, followed by an optional Vienna RNA display.
- The edge list is a triangular matrix showing Include (i) or Overlapping (o) relationships (edges) between stems (vertices).

- The adjacency matrix is a square matrix showing the relationships, X, I, O, and S, between all of the stems.

2.2 Motif library generation: enumerating a comprehensive set of RNA topologies

We have developed a ***structural motif library***, which is an exhaustive enumeration of all possible RNA structural motifs. Each motif, a graph with a fixed number of stems, is represented by a XIOS graph and assigned a canonical DFS code. The current structural motif library contains 55,728 motifs in total, which represents all physically non-redundant motifs containing from 1 to 7 stems (Table 2.1). The motifs in the library contains either I, O, and S edges. The set of motifs with N stems is generated by generation of all the permutations of an ordered set of 2N numbers, and removal of redundant graphs with isomorphism. Two graphs are considered to be isomorphic if they have the same number of vertices and the ways of the vertices being connected are the same. Graph isomorphism is important because we can use it to identify identical graphs and thus identify similar RNA functions. Graph isomorphism is identified by canonical labeling termed minimum ***depth-first search*** (***DFS***) code using the gSpan (104,107) approach. For further details on the motif library generation, refer to Chapter 3.

2.3 RNA graph matching using DFS lexicographical ordering

With the XIOS graphical representation, one is able to compare RNA structures based on topology; however, graph matching is an NP-complete problem. Thus, an efficient graph

enumeration and matching technique is required. Inspired by the *DFS* lexicographical ordering (107) of gSpan, we have modified this approach to match XIOS graphs.

In the gSpan approach, a graph can be canonically represented by enumerating its vertices and edges following lexicographical rules. This canonical enumeration is the DFS code. Two graphs with the same DFS code are isomorphous (107). Figure 2.3 shows an example of the DFS code for a XIOS RNA graph.

2.4 RNA fingerprint generation and XPT format

2.4.1 RNA fingerprint and its generation

Structural motifs are an intrinsic property of an RNA structure. We define the ***RNA structural fingerprint*** (or simply, ***fingerprint***), as a list of the structural motifs found in a specific RNA structure. In terms of the XIOS graph, the fingerprint is a list of its subgraphs. Figure 2.4 shows an example of RNA fingerprint.

Figure 2.5 shows a flow chart of RNA fingerprint generation for a RNA XIOS graph. We have developed a subgraph random sampling algorithm that identifies the subgraphs in a XIOS graph. The identified subgraphs are then encoded as DFS codes to allow matching of motifs with the same DFS code. When a certain number of iterations is reached or specific conditions are satisfied, the fingerprint generation stops (Chapter 3.3.2), and a list of identified motifs is written into a fingerprint file (the XPT file).

2.4.2 XPT format

Similar to the XIOS files, we have defined XPT files, which are also written in XML format. Figure 2.6 is an example of the XPT file of RNase P RNA from *A.fulgidus*. An XPT file contains four blocks: query, fingerprint, database, and motif list. The query block contains the RNA graph id, and the number of its vertices and edges. The fingerprint block shows the statistics of the fingerprint computation, including iterations, run time, and the version of the program used. The database block shows the id of the motif library used. The motif list shows each motif (subgraph) identified in the XIOS graph: the motif id, corresponding DFS code, the first iteration of iterations of the motif being sampled in random subgraph sampling (Chapter 3), total number of iterations of the motif being sampled, and total number of different mappings (one mapping is one combination of vertices composing the motif) of the motif in the RNA graph. The DFS code is rewritten as a hexadecimal code for simplification.

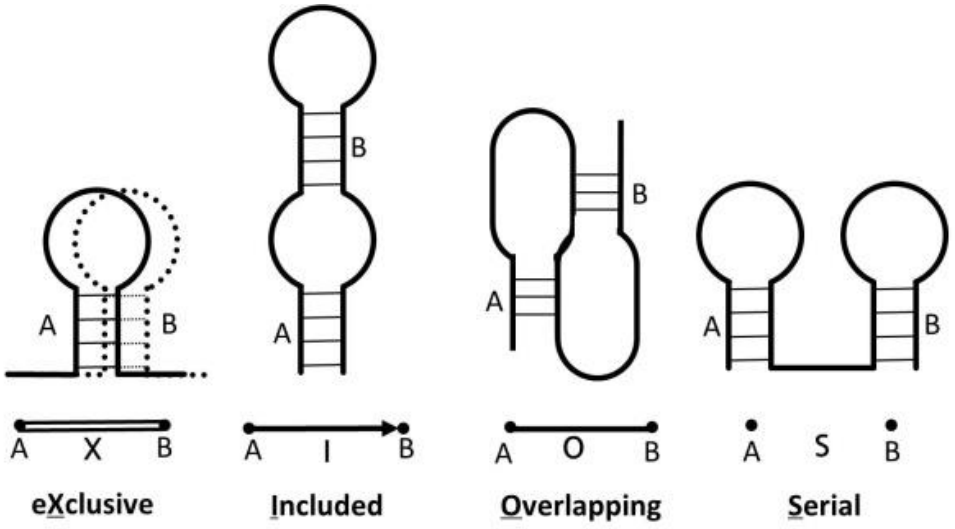


Figure 2.1 XIOS graph stem-stem relationships. Edges show the relationship between two stems, and may be one of four types: X (mutually exclusive), I (included or nested), O (overlapping or pseudoknotted), or S (serial or adjacent). The bold lines indicate the sequence, and the thin lines indicate the base-pairing between specific regions on the sequence. A and B are two stems. In the left-most panel, the dotted lines indicate the alternative base-pairings that form stem B, other the base-pairing in solid lines that form stem A.

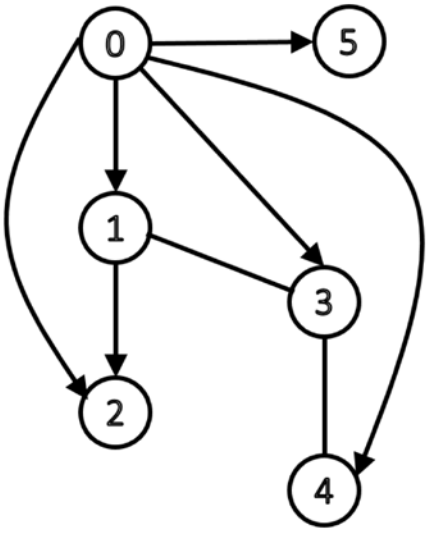
```
<XIOS>
      <information>
            <sequence>
                  <id>A.fulgidus</id>
                  <doc>RNase P RNA</doc>
                  <unformatted>CGGCGGUGGGCGGCUGACCGAAAGGAGGAAAGUCCCCCCACCC
                  GCUGUGGCGGAAGGCCCCUGAGAAGGGGCGGAGGAGGAACAGAAACGAGACCGGUG
                  CGGGGAAAUGCGAUGAUUCCGCAAGGAUGAGGUCACCCGCUCCGGAUGAAACGGCC
                  UCCUCCCCGCCGGGUGCAACGCGUAAGCGGCUCAGUCUAAUGCCGCCGGAACAGAA
                  GGGGGCUUACUACCGCCA</unformatted>
            </sequence>
      </information>

      <stem_list>
           0 115.0 [   1   8   222 229 ]      ((((((((   ))))))))
           1 105.5 [   9  14   197 202 ]        ((((((   ))))))
           2  21.5 [  18  19    24  25 ]            ((   ))
           3 125.0 [  30  38   212 219 ]     (((.(((((   ))))))))
           4 105.0 [  39  43   167 171 ]         (((((   )))))
           5 107.5 [  49  53   162 166 ]         (((((   )))))
           6  64.5 [  57  62    67  72 ]        ((((((   ))))))
           7 117.0 [  73  81   153 161 ]     (((((((((   )))))))))
           8 118.5 [  94 105   132 144 ]  ((((((((((((   )).))))))))))
           9 121.5 [ 117 119   124 126 ]           (((   )))
          10 179.5 [ 175 177   182 184 ]           (((   )))
      </stem_list>

      <edge_list>
           0:  1i  2i  3i  4i  5i  6i  7i  8i  9i 10i
           1:  2i  3o  4i  5i  6i  7i  8i  9i 10i
           2:
           3:  4i  5i  6i  7i  8i  9i 10i
           4:  5i  6i  7i  8i  9i
           5:  6i  7i  8i  9i
           6:
           7:  8i  9i
           8:  9i
           9:
          10:
      </edge_list>

      <adjacency>
              0  1  2  3  4  5  6  7  8  9 10
           0  -  i  i  i  i  i  i  i  i  i  i
           1  j  -  i  o  i  i  i  i  i  i  i
           2  j  j  -  s  s  s  s  s  s  s  s
           3  j  o  s  -  i  i  i  i  i  i  i
           4  j  j  s  j  -  i  i  i  i  i  s
           5  j  j  s  j  j  -  i  i  i  i  s
           6  j  j  s  j  j  j  -  s  s  s  s
           7  j  j  s  j  j  j  s  -  i  i  s
           8  j  j  s  j  j  j  s  j  -  i  s
           9  j  j  s  j  j  j  s  j  j  -  s
          10  j  j  s  j  s  s  s  s  s  s  -
      </adjacency>
</XIOS>
```

Figure 2.2 XIOS format.



v_i	v_j	$e_{i,j}$
0	1	I
1	2	I
2	0	J
1	3	O
3	0	J
3	4	O
4	0	J
0	5	I

Figure 2.3 An example of the conversion of a XIOS graph into a canonical DFS code representation. In a DFS tree, an edge is represented by a 3-tuple, (v_i, v_j, $e_{i,j}$); v_i and v_j represent two stems connected by edge $e_{i,j}$, which is X, I, O, or S in an XIOS RNA graph. According to the DFS lexicographical ordering, a DFS tree is extended by adding edges as follows: edges connecting to previously identified vertices (backwards edges), edges connecting the most recently added vertex to a new vertex (forward edges), and edges connecting internal vertices to a new vertex. If two possible edges can be extended from the same vertex, first enumerate I edges, then J edges, and in the end, O edges. The left panel shows the canonical DFS tree, which can be represented by the DFS code in the right panel.

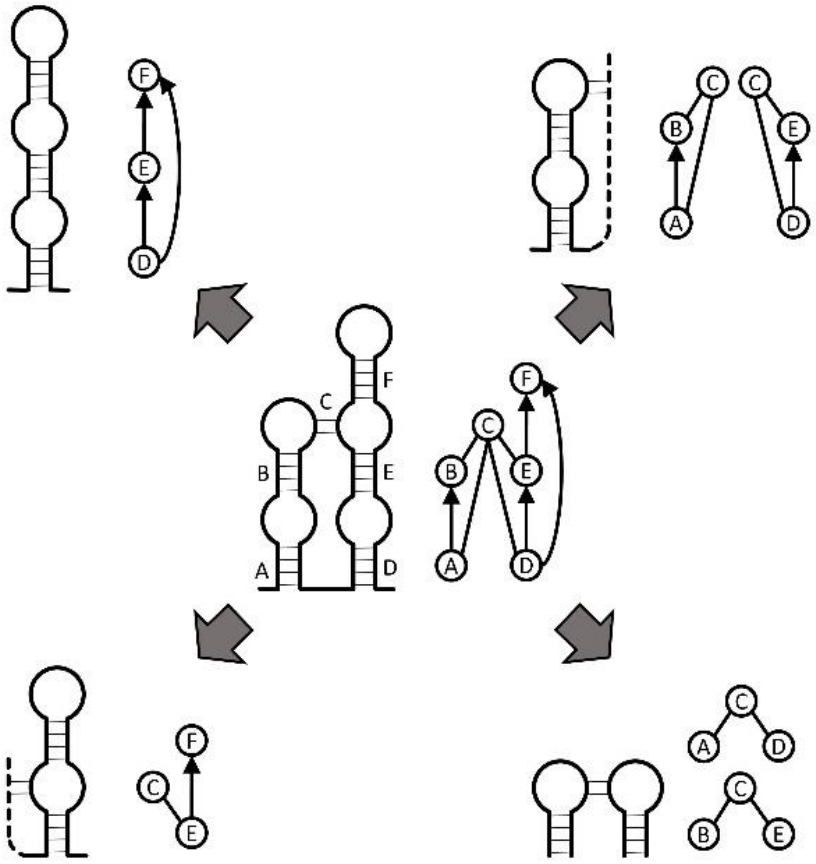


Figure 2.4 Example of an RNA fingerprint. All 3-vertex subgraphs (corners) in a 6-vertex RNA graph (center) are shown. The subgraphs comprise the 3-fingerprint of the RNA graph.

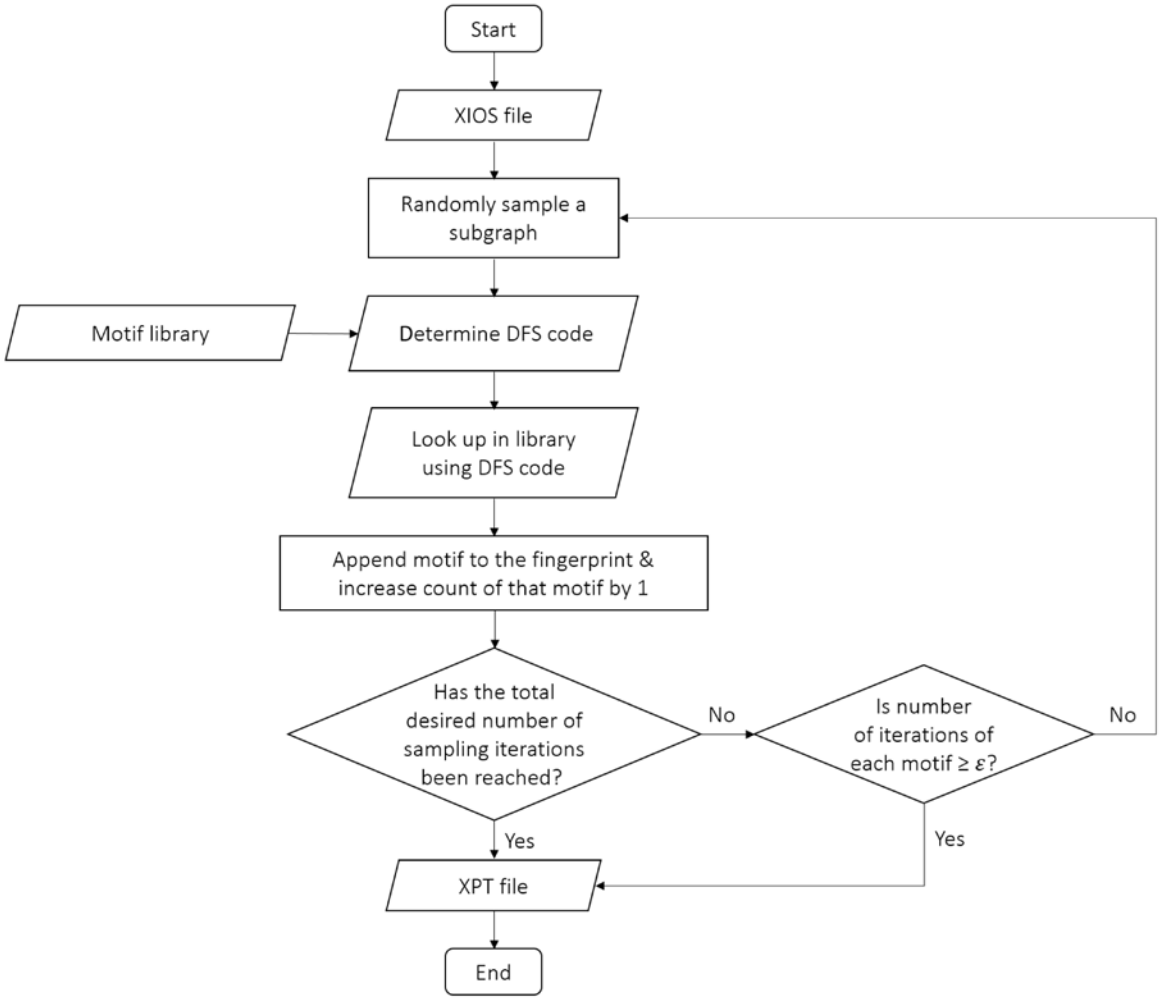


Figure 2.5 Flow chart of fingerprint generation. Ɛ is the threshold of iterations of each sampled motif, arbitrarily set to be 10 in this study (Chapter 3.3.2).

```
<XIOS_fingerprint>

       <query>
              <query_id>rnasep_m.A_fulgidus.xios</query_id>
              <query_vertex>11</query_vertex>
              <query_edge>38</query_edge>
       </query>

       <fingerprint>
              <iteration>10000</iteration>
              <program>fingerprint_random.pl v1.1.2.16</program>
              <time_elapsed>25.422513</time_elapsed>
       </fingerprint>

       <database>
              <database_id>2_to_7_stems_topologies.removed_not_true.mini_dfs.
              txt.removed_redundant.with_label.motif.storable</database_id>
       </database>

       <motif_list>
              <motif_n>4</motif_n>
              <motif>
                    <id>7_11091</id>
                    <count>634</count>
                    <first_observed>42</first_observed>
                    <encoded_dfs>0428414c61657081858954a1a538c1</encoded_dfs>
                    <mapping>18</mapping>
               </motif>
               <motif>
                    <id>7_30045</id>
                    <count>628</count>
                    <first_observed>6</first_observed>
                    <encoded_dfs>0428414c61657081858954a1a518</encoded_dfs>
                    <mapping>21</mapping>
               </motif>
               <motif>
                    <id>7_36615</id>
                    <count>581</count>
                    <first_observed>14</first_observed>

       <encoded_dfs>0428414c61657081858994a1a5a9ad38c1</encoded_dfs>
                    <mapping>15</mapping>
               </motif>
               <motif>
                    <id>7_35341</id>
                    <count>565</count>
                    <first_observed>23</first_observed>
                    <encoded_dfs>0428414c61657081858934a118</encoded_dfs>
                    <mapping>22</mapping>
               </motif>
           </motif_list>
</XIOS_fingerprint>
```

Figure 2.6 XPT format. The DFS code is rewritten as a hexadecimal code for simplification.

Table 2.1 Topological Motif Library.

Number of Stems	Unique Topologies
1	1
2	2
3	8
4	46
5	368
6	3,914
7	51,390
total	55,728

CHAPTER 3. ACCURATE CLASSIFICATION OF RNA STRUCTURES USING TOPOLOGICAL FINGERPRINTS

While RNAs are well known to possess complex structures, functionally similar RNAs often have little sequence similarity. While the exact size and spacing of base-paired regions vary, functionally similar RNAs have pronounced similarity in the arrangement, or topology, of base-paired stems. Furthermore, predicted RNA structures often lack predicted pseudoknots (a crucial aspect of biological activity), and are only partially correct or incomplete. A topological approach addresses all of these difficulties. In this work we describe each RNA structure as a graph that can be converted to a topological spectrum (RNA fingerprint). The set of subgraphs in an RNA structure, its RNA fingerprint, can be compared with the fingerprints of other RNA structures to identify and correctly classify functionally related RNAs. Topologically similar RNAs can be identified even when a large fraction, up to 30%, of the stems are omitted, indicating that highly accurate structures are not necessary. We investigate the performance of the RNA fingerprint approach on a set of eight highly curated RNA families, with di-verse sizes and functions, containing pseudoknots, and with little sequence similarity – an especially difficult test set. In spite of the difficult test set, the RNA fingerprint approach is very successful (AUC > 0.95). Due to the inclusion of pseudoknots, the RNA fingerprint approach both covers a wider range of possible structures than methods based only on secondary structure,

and its tolerance for incomplete structures suggests that it can be applied even to predicted structures.

3.1 Introduction

Once seen as a simple scaffold, RNA is now known to play important regulatory and catalytic roles. RNA is involved in processes including transcriptional regulation (108), RNA maturation and modification (8), and RNA splicing (10). The structural motifs in RNA that are responsible for its functions are evolutionarily conserved; however, unlike DNA and protein, for which conserved functional motifs can be identified based on sequence similarity, the functional motifs in RNA may have little or no sequence similarity (109), and instead conserve patterns of base-pairing (stems) and topological relationships between base-paired regions, for instance nesting of stems, multi-loops, and pseudoknots (26,110). This topological view of RNA structure has been discussed by Giegerich et al. who point out that, in a family of RNAs with the same function, the global arrangements of structural elements (topology) are conserved, but there is considerable variation in the length of stems, presence of bulge loops and unpaired bases, and type of base-pairs. Therefore, in the study of RNA functions, it may be more relevant to look at global topological patterns than individual base-pairs (94,95). RNAs with similar functions, for example those in ribonuclease P (RNase P), the ribosome, or self-splicing introns, typically have strongly conserved topologies (15,26,28,111). One of the notable topological aspects of RNA structure is the importance of pseudoknots in many classes of molecules. For example, in Hepatitis Delta Virus (HDV), a double-pseudoknotted structure contained in a self-cleaving ribozyme is a key factor in HDV infection (112); in Group I self-

splicing introns, the catalytic core is formed by pseudoknots (29); in ribosomal RNA, pseudoknots at the catalytic site are the key structures that mediate microbial resistance to antibiotics (113) and stimulate viral frame-shifting (114).

As only a small number of functional RNA classes have been identified, we believe that the majority of regulatory and functional RNA motifs are yet to be identified. Eukaryotic genomes are pervasively transcribed (1); almost every base can be found in an RNA transcript. This is surprising since, in most genomes, protein-coding sequences comprise only a small fraction of the genome. Much of this RNA is therefore likely to be regulatory in nature, and will almost certainly contain functionally important structures, including pseudoknots.

Just as conserved structural topologies are important for RNA function, the identification of novel conserved topologies provides an approach to discovering the functions of currently unknown classes of biologically important RNAs. An analogy can be made to the importance of sequence alignment and database searching programs in identifying novel proteins and DNA regulatory elements. While typical functional RNA structures are pseudoknotted, the current computational approaches to RNA structure comparison only consider structures without pseudoknots. Because of their importance to RNA function, we believe that incorporating pseudoknots in structural comparisons is critical to identifying biologically important classes of molecules. In this paper we propose a straightforward approach to comparing RNA structural topologies, including pseudoknots, and identifying known and unknown conserved topologies.

Waterman (115) introduced the first graphical representation of RNA structure, the tree-graph. The tree-graph representation was extended by Shapiro et al. to an abstract tree where the nodes represent structural elements (90-92), and this coarse-grained representation was implemented in the ViennaRNA package (89). Fontana et al. implemented the homeomorphically irreducible tree (HIT) that represents an RNA secondary structure as a contracted topology in which each node represents a structural element weighted by size (93). Shu et al. have developed the element-contact graphs (ECGs) with size-weighted nodes as well (116), which uses topological indices, such as the Randić index(116,117), the Wiener index, and Balaban index, to measure graph connectivity. Although the ECGs framework was shown to be able to identify small ncRNAs such as miRNAs, no evidence is shown for its ability to classify larger RNAs (for example, 23S rRNA are usually over 1000nt long) with low sequence similarity. The RNAshapes package (94,95) of Giegerich et al., which represents RNA structures as abstract shapes and aims for efficient RNA structure comparisons, has been shown useful in topologically clustering RNA families; however, RNAshapes does not perform well on families with pseudoknots (96). Building on this work, Heyne et al. developed a graph-based pipeline called GraphClust (118) for fast clustering of RNA molecules. In this approach, RNA secondary structures are generated by the RNAshapes package from input sequences, encoded by graphs preserving nucleotide connectivity, and clustered by a graph kernel, the Neighborhood Subgraph Pairwise Distance Kernel (NSPDK) (119). However, given data sets of small RNA sequences (sequence length < 400nt, similarity up to 80%) the

precision and recall of GraphClust only reaches around 85%. In addition, these approaches do not include pseudoknots in either the representation or the analysis.

The Schlick group has developed the RNA-As-Graphs method, which represents RNA structures as tree graphs, without pseudoknots, or dual graphs, with pseudoknots (97-100). Numerical descriptors have been applied to comparison of these RNA topological patterns. The eigenvalue spectrum of the Laplacian matrix measures graph compactness and connectivity; λ_2, the second eigenvalue of the Laplacian matrix (120,121), measures RNA graph similarity. The Schlick group used several structural invariants, including λ_2 and linear combinations of α and β (the intercept and slope of the eigenvalues of the Laplacian matrix), for categorizing the structural similarity of RNA graphs, and for predicting whether randomly generated RNA topologies are similar to biological examples (RNA-like). These numerical descriptors, however, have never been shown to be able to group RNAs into structural/functional classes. Moreover, these approaches, which rely on a small number of numeric descriptors, cannot identify similarity between specific substructures nested within fairly large graphs (for instance graphs of the size of RNase P RNA, which may have up to 20 vertices).

There are several aspects of RNA structure that make it particularly hard to identify topologically similar structures. Structures from the same functional family may have little or no sequence similarity; they typically have a similar arrangement of stems (topology), but different local base-pairing; our knowledge of the structures may be incomplete due to lack of a high-quality three-dimensional structure or structural prediction; structures may lack biologically important pseudoknots since tractable computational approaches

based on dynamic programming often do not include these important features; or in the case of graph comparison, the computation itself may require infeasible amounts of time. The RNA XIOS graph (104) explicitly represents serial, nested, pseudoknotted, and mutually exclusive stems, but finding topologically similar RNA structures requires identifying isomorphous subgraphs common to one or more structures. The approach we describe here builds on the XIOS approach, addresses the problems described above, and provides a feasible approach to identifying biological RNAs with topologically similar structures. We demonstrate the utility of this approach by classifying a representative set of pseudoknot-including RNA structural families that have very low levels of sequence similarity – the high accuracy of the classification indicates that this approach can be broadly applied to identifying RNAs with conserved topologies, whether their function is known or unknown.

3.2 Materials and methods

3.2.1 Curated RNA families

A set of curated RNA structures have been collected from the literature and a variety of biological databases (106) and is extended in this work (Table 3.3). This set of known structures has been carefully selected to contain pseudoknots, to cover a broad range of lengths, and to have been the subject of extensive expert curation by the biological community. This curated set includes 206 structures of transfer RNA, Ribonuclease P RNA, transfer-messenger RNA, group I and group II self-splicing introns, and 5S, 16S and 23S ribosomal RNA. The structures in this curated set have been reviewed to ensure

they reflect expert opinion on the correct structure, and to ensure that the reported structures are as accurate as possible given existing experimental data such as X-ray crystallography (122,123) and covariance analysis (85). The curated structures have been screened to ensure that all structures are full-length, and no pair of structures has greater than 50% sequence identity. Multiple families of the curated structures contain pseudoknots. While several large databases of RNA structures exist, for instance Rfam (124) and RNAStrand (125), these databases suffer from a number of disadvantages that make them difficult to use as a gold standard. Among the problems in these extensive datasets are the lack of pseudoknots in many structures, a lack of consensus expert opinion on the correct structures, the presence of families for which only a family consensus structure is available (rather than individual structures for each RNA), high levels of sequence identity within families, and the presence of incomplete structures, or structures in which single stranded regions (or other regions judged to be unimportant) have been removed.

3.2.2 XIOS graphs

In a XIOS graph, RNA stems are shown as vertices and the relationships between stems are shown as edges (104). Edges may be one of four types: X – mutually exclusive (stems with base conflicts, such as those in two alternative structures that use the same RNA sequence); I – included (nested); O – overlapping (pseudoknotted); S – serial (adjacent) (Figure 2.1). Because there are exactly four classes, and each pair of stems can have one and only one type of relationship, we can omit S relationships without loss of generality

(any pair of vertices without an edge have an implicit S edge). In this work, none of the structures have X edges; the graphs therefore have only two edge types, I and O. Figure 3.5 shows the XIOS graph representation of the Hepatitis D Virus (HDV) ribozyme RNA.

3.2.3 Curated XIOS graphs

Table 3.3 shows the vertex number, edge number, and average degree of the XIOS graphs of the curated RNA structures. Graph matching is highly dependent on the size of the graph (described by the number of vertices and edges) and the average degree of the vertices in the graph; the characteristics of the curated RNA structures differ significantly between families making this a representative set for RNAs in general.

3.3 Results

This work focuses on the topological similarity between RNA structures, that is, similarity in the relative location and nesting of stems, and the location of pseudoknots. In principle, this should provide the broadest range of matching since individual structures often differ in the length of stems and the length of single-stranded regions between stems. As mentioned before, the sequences themselves can be even more variable with little or no sequence conservation detectable, even between RNAs with similar structures. Topologically similar substructures in a pair of RNAs correspond to isomorphous subgraphs in their respective XIOS graphs. The maximal common subgraph (MCS) represents the greatest possible topological match between RNAs, similar to the maximal alignment between two sequences. But the MCS is difficult to identify because of the

large size of biologically important structures; e.g., the 23S rRNA can have more than 50 stems (126). Finding the MCS of a set of graphs, corresponding to the largest conserved topological motif in a group of RNA structures, is an NP-hard problem (127), making the computational identification of the MCS time consuming. In order to decrease the inefficient scaling inherent in graph matching, we characterize each graph as a set of smaller subgraphs. We call this set of subgraphs the RNA topological fingerprint, or more simply, the RNA fingerprint. There are two key elements needed to determine an RNA fingerprint: a comprehensive dictionary of RNA topological motifs, and an approach to identifying the motifs that are present in a XIOS graph.

3.3.1 Enumerating a comprehensive set of RNA topologies

We have exhaustively enumerated a non-redundant set of all physically possible RNA topological motifs containing from one to seven stems (Table 2.1). The graphs in this set are all IO-connected, that is, all vertices (stems) can be reached by traversing I and O edges. Briefly, a complete set of topologies for an N-stem RNA structure can be created by generating all the permutations of an ordered set of 2N numbers; the numbers represent N objects (stems), numbered 1 to N, each with two instances (corresponding to the two base-paired halves of the stem). For three stems (N=3), the ordered unpermuted set would be (1, 1, 2, 2, 3, 3), with each pair of matching numbers representing the two base-paired halves of a stem. The unpermuted set, above, would thus correspond to three serial stems, and a permuted set such as (1, 2, 3, 2, 3, 1) would indicate a pair of pseudoknotted stems, 2 and 3, found within the loop of stem 1.

Obviously, this procedure generates multiple copies (isomorphs) of some topologies, for instance (1, 2, 2, 3, 3, 1) and (3, 1, 1, 2, 2, 3), as well as some graphs that are not connected (for instance the unpermuted set, above). Some of the isomorphs can be eliminated by imposing two restrictions. First, the graph must be connected, and second, the first instances (left half stem) of each object (stem) must occur in numerical order. Even these restrictions do not entirely eliminate permutations that correspond to isomorphic XIOS graphs. For instance, the sets (1, 2, 1, 3, 3, 2) and (1, 2, 2, 3, 1, 3) are mirror images of each other, and correspond to the same XIOS graph. These symmetry-related topologies are detected and removed using the gSpan (104,107) approach. In gSpan, a graph is described using a canonical labeling called the minimum DFS code; Isomorphic graphs are guaranteed to have identical minimum DFS codes.

Using this approach, we have enumerated a library of all unique physically possible RNA topologies with 2 to 7 stem structures (Table 2.1). Because the minimum DFS code provides a unique description for each topology, we index the motif library with a compressed version of the minimum DFS code. The index of any structure within the library can be easily determined by simply determining its minimum DFS code.

The topologies in the library are not independent; two unique 5-stem XIOS graphs, for instance, may share a common 4-stem subgraph as shown in Figure 3.1. In this situation, we say that the 4-stem subgraph is the parent of both 5-stem graphs because they each have had one stem added to the parent subgraph (Figure 3.1). When comparing topological motifs, subgraphs that share a parent are clearly more similar than subgraphs that only share a grandparent or great-grandparent. The topological motif library in-

cludes all the parent and child relationships between the enumerated graphs in order to allow for partial matching.

3.3.2 Determining RNA fingerprints using random sampling

A XIOS graph corresponding to a single structure can be characterized by the set of fixed-size subgraphs it contains. This set of constituent subgraphs is the RNA fingerprint (Figure 2.4), which can be thought of as a subgraph spectrum that is characteristic of a specific topology. Currently we use a library comprising all 7-stem and smaller subgraphs; this number has been chosen to cover both large and small biological structures, without requiring excessive computation. For even a relatively small graph, for instance a graph with 25 to 30 vertices, exhaustively enumerating the complete set of 7-vertex subgraphs within it can be time consuming. The subgraph sampling approach we describe here allows the determination of the fingerprint in reasonable time on parallel hardware. Briefly, given a XIOS graph, we randomly sample a fixed number, currently seven, of connected vertices from the graph (Table 3.1). Sampling continues until a suitable termination condition is met, typically when all observed subgraphs have been independently sampled 10 times. In each iteration, one subgraph is sampled and uniquely identified by its minimum DFS code, which is used as a reference to identify the subgraph in the RNA structural motif library. The complete fingerprints of 151 RNA structures computed by an exhaustive method (not shown) have been used to validate the correctness of the RNA fingerprints computed by random sampling (Figure 3.2).

3.3.3 RNA fingerprints identify topologically similar RNA structures

The set of subgraph motifs sampled in a query graph is its simple fingerprint. We define the extended fingerprint as the simple fingerprint plus all of the ancestral subgraphs (i.e., parent, grandparent, etc., see Figure 3.1) of the simple fingerprint motifs. In this section we use both the simple fingerprint and the extended fingerprint to identify RNAs with similar topologies. The average numbers of motifs in simple and extended fingerprints are shown in Figure 3.6.

Consider the simple or extended fingerprints, X and Y, of RNA R_X and RNA R_Y; X = $\{x_1, x_2, x_3, ..., x_m\}$ and Y = $\{y_1, y_2, y_3, ..., y_n\}$ where $x_1, x_2, x_3, ..., x_m$ and $y_1, y_2, y_3, ..., y_n$ are the subgraph motifs found in RNAs R_X and R_Y. We have evaluated five similarity functions (Table 3.2) for their ability to identify topologically similar structures.

Figure 3.3 shows the classification performance of the different similarity functions as measured by Receiver Operating Characteristic (ROC) curves (128). Jaccard Similarity works best in the classification of RNA structures, with an area under the ROC curve (AUC) greater than 0.95 for the extended fingerprint. The increase in AUC from 0.870 for the simple fingerprint to 0.952 for the extended fingerprint using Jaccard Similarity indicates that the inclusion of parent subgraphs substantially improves the detection of topologically similar structures. The classification performance of Jaccard Similarity using the extended fingerprint on different RNA classes is around 0.95 for all groups except for 16S rRNA and group II introns (Table 3.4). Figure 3.4 shows the ability of the extended-Jaccard similarity to effectively classify the test structures into functional groups. As can be seen in the upper triangle of Figure 3.4, the level of sequence similarity is very

low between these structures and would be insufficient for correct clustering (not shown). The 23S rRNAs form a single group, and also share some similarity with 16S rRNAs, which may be explained by the topological similarity of the two subunits of rRNA (126). The 5S rRNAs form two separate groups, one with archaeal and eukaryotic nuclear structures, and the other with bacterial structures. Self-splicing introns, especially the Group II Introns, share a high topological similarity with the 23S and 16S rRNAs. The accuracy of the classification confirms that our topological approach can identify topologically similar RNAs, and potentially functionally similar RNAs, as well. In addition, a neighbor-joining tree (129) (Figures 3.4, 3.7 and 3.8), using the extended-Jaccard similarity, correctly groups almost all the curated RNA families into the correct categories, with only one Group I Intron falling onto a branch outside of its curated group (Figure 3.4, tree on the right side).

3.3.4 Similarity of incomplete graphs can be detected using RNA fingerprints

In most cases, topological comparisons must be based on predicted structures, because three-dimensional structures or high-quality comparative structures are usually unavailable. Although structures with pseudoknots can be predicted (75,81,106,130,131), such predicted structures will typically be inaccurate or incomplete. It is highly desirable that a similarity function be able to correctly identify similar RNAs, even when their structures are incomplete. To test the effects of graph incompleteness on the extended-fingerprint Jaccard Similarity function, incomplete RNA graphs were generated by randomly removing a percentage (10%, 30%, 50%, 60%, and 70%, respectively) of the verti-

ces (stems) in the curated structures (Figure 3.3F). The extended-fingerprint Jaccard Similarity can identify similar structures when only 70% of the original stems are present (AUC=0.810), and performs better than random even when only 30% of the stems remain. In addition, since pseudoknots are important structural motifs in RNAs, for the 149 RNA structures that have pseudoknots, we generated incomplete RNA graphs by first removing all the pseudoknot-forming vertices (stems), and continuing removing random vertices until 30% of vertices were removed. The extended-fingerprint Jaccard Similarity correctly identifies similar structures with pseudoknots removed (AUC=0.915, data not shown).

3.3.5 Fingerprint similarity is not an artifact of graph size

The structures within each curated family generally have very similar numbers of stems. Indeed, one can classify the structures into the correct groups using graph size alone (not shown). It is essential, therefore to consider whether the results in Figures 3.3 and 3.4 are merely due to the similarity in sizes. In order to test the effect of size, we have created a test data set in which the graphs have been expanded to the same size (number of vertices) by randomly adding additional vertices and edges to the graphs. In order to ensure that these expanded graphs are typical of real biological structures we use a procedure in which we sample substructures from the set of curated structures, and add them to the curated graphs. In order to do this, we created a database (decoy database) of the 2 to 5 stem motifs found in the curated structures, and randomly added these

subgraphs to the curated structures according their frequency in the entire curated set (which should reflect the biological background distribution).

We selected a set of 177 RNA graphs containing up to 25 vertices from the curated data set (Table 3.3), and created an expanded set by embedding subgraphs, randomly selected according to probability of occurrence, from the decoy database into these RNA graphs until each RNA graph contained 30 vertices. As a control, we also created a decoy set of graphs with 30 vertices, by random embedding of subgraphs from the decoy database only, *i.e.*, graphs with no information from real biological structures except the frequency of occurrence of subgraphs in the known structures. Both the expanded and the decoy graph sets should be completely free of size effects since they all have exactly the same number of stems. The two sets were mixed and graphs compared using the Extended Fingerprint Jaccard Similarity. There is only a minor decrease in performance (Table 3.5, Extended Fingerprint Jaccard Similarity: AUC = 0.840) when compared to the results obtained from the classification of the original dataset (Figure 3.3, Extended Fingerprint Jaccard Similarity: AUC = 0.952). As expected, the decoy set of graphs have AUC values close to 0.5, indicating that the decoy structures are random with respect to each other.

3.3.6 Runtime analysis

Determination of whether a query RNA graph contains a subgraph isomorphic to a specific graph in the structural motif library, is an NP-complete problem (127). The brute-force comparison requires comparing the query RNA graph with every graph in the li-

brary, and its computational complexity is $O(nm^m)$, where *n* is the number of graphs in the library (55,728), and *m* is the number of edges in the query graph. The subgraph random sampling algorithm can be parallelized by simultaneously running independent instances on multiple processors. The algorithm identifies the fingerprint of all 206 curated RNA graphs in a reasonable time, especially when it is run on multiple cores (Figure 3.9). The average runtime for calculating the fingerprint of RNAs in each functional family is shown in Table 3.6.

3.4 Discussion

A great deal of work has focused on identifying similar RNAs based on the comparison of RNA secondary structures. This is readily accomplished using approaches such as tree edit distance (93,132) or string related measures such as those used in RNAshapes (94). Other approaches include the information of sequence alignment and folding of RNA sequences, for example, Saito *et al.* developed an algorithm that clusters RNAs by all possible sequence alignments, and all possible secondary structures computed from dynamic programming and partition function calculations (55,133,134). This approach correctly discriminated short RNA sequences (around 100 bases) from different families. Unfortunately, secondary structures, and in particular minimum free energy predicted structures based on dynamic programming approaches, do not predict pseudoknots, which are important in biological structures. Even if predicted pseudoknots are available, it is not simple to add them to tree or string based methods because of their non-nested nature. In addition, structure matching methods based on dynamic programming

have the additional problem of determining gap penalties; it is not at all clear how to weight insertions and deletions in RNA structures.

Statistical algorithms, such as kernel methods, have been developed to classify RNA sequences and structures. Kin et developed a marginalized kernel to measure RNA sequence similarity (135), and this kernel was later implemented by Karklin *et al.* to measure the similarity of RNA secondary structures represented by dual graphs (97,136); Liu *et al.* developed a fuzzy kernel to cluster the secondary structure ensemble generated from a single sequence (137). The GraphClust pipeline developed by Heyne *et al.* encodea RNA sequence-structure information into graphs and measures RNA graph similarities using a decomposition kernel and computing the summed similarity of pairs of neighborhood subgraphs (138). However, no pseudoknotted structures were included in these approaches. Sakakibara *et al.* developed a stem kernel that could discriminate between functional RNA sequences and randomly shuffled sequences using structural features including pseudoknots (139); however, no result was shown in which the stem kernel could discriminate between sequences from different functional RNA groups, in addition, the randomly shuffled sequences they generated only retain nucleotide composition, while preserving dinucleotide composition is known to be important in generating randomized negative controls for predicted RNA structures (140,141). In summary, none of these approaches have demonstrated that they can succeed on the difficult test case presented here: classifying a diverse set of functional families, with diverse sizes, containing pseudoknots, and with little sequence similarity.

Topological methods have the dual advantage of easily representing pseudoknots and not requiring an insertion/deletion penalty. In the RNA-As-Graphs method (97,99,100,142), RNA topologies are represented either as "tree graphs" (without pseudoknots) or "dual graphs" (with pseudoknots). Gan *et al.* suggested summarizing the topological properties of an RNA graph using the eigenvalue of the Laplacian matrix (constructed from the adjacency and degree matrices of the graph). They have developed a database, with all mathematically possible RNA graphs enumerated, including "existing graphs" (RNA structures experimentally solved or obtained from comparative analysis) and "missing graphs" (mathematically possible RNA structures that have not yet been experimentally observed). Using "existing graphs" as training data, "missing graphs" in the database were classified as either "RNA-like" or "non-RNA-like" by applying regression analysis on their Laplacian eigenvalue spectra. These approaches, which target the identification of novel RNA topologies, however, are not sufficient for matching specific RNA functional families.

Graph matching is a computationally intensive process that scales exponentially with the size of graph (in general, graph matching is an NP-hard process) (143). The RAG database, however, only includes dual graphs up to 9 vertices and tree graphs up to 10 vertices (142), which can cover RNA topologies only up to about 200nt, while functional RNA molecules can include dozens of stems/loops, especially with the current advance in high-throughput technologies, and long non-coding RNAs including hundreds of stems/loops are not uncommon (144). Moreover, in a follow-up study, the discrimination between structures predicted to be RNA-like (naturally existing) and not

non-RNA-like was not impressive; out of 42 newly discovered RNA topologies, only 24 of them had been predicted as RNA-like, while 18 of them had been predicted to be non-RNA like (142).

The XIOS graph is a topological graph approach (104) that specifically distinguishes pseudoknots as a distinct type of edge. In addition to incorporating pseudoknots (O edge, Overlapping), one of the most important characteristics in RNA structure, the XIOS approach also includes embedding (I edge, Included) and juxtaposition (S edge, Serial), which are the two of the RNA structural principles in the RNAshapes framework. The increased number of edge-types in XIOS improves one's ability to match graphs, for example using gSpan; however, the time required to find the maximal common subgraph in two moderately large RNA graphs, for instance with twenty to thirty stems in each graph, is prohibitive using exhaustive approaches such as gSpan. Using the XIOS approach, we can easily enumerate a complete set of biologically possible RNA graphs, permitting the construction of a complete dictionary of all graphs that may occur in a RNA molecule up to a specified size. This allows us to characterize any RNA topology in terms of the spectrum of subgraphs it contains, its RNA fingerprint, and to identify topologically similar RNA structures based on their fingerprints. This approach is successful with known RNA families, and is relatively insensitive to both the completeness of the RNA graph, and the presence of extraneous added vertices in the graph. Similarities between RNA structures in the same family are still detectable when the graphs are expanded to the same size, indicating that the ability to identify topologically similar structures is not simply an artifact of the similar sizes of RNAs within known families. These

characteristics of RNA fingerprint matching are highly important in real-world settings where comparisons are made between predicted structures in which only 60 - 80% of the true stems may be correctly predicted (131,145), and a substantial number of mis-predicted stems may be present. As mentioned before, no previously reported RNA structure comparison method has shown that it can accurately identify/classify RNAs according to topological similarity using the particularly difficult set of pseudoknot containing graphs used here.

Exhaustively enumerating the set of subgraphs present in a XIOS graph is time consuming because each subgraph in the entire motif database must be separately tested against the query to determine whether there is a match. Because the dictionary of subgraphs is large (55,728 graphs with seven or fewer stems), a brute force approach is slow. In this work we suggest a sampling approach to enumerating the subgraph spectrum. The computational complexity of motif sampling depends on both the size and structure of the query graph, and on the number of vertices sampled in each iteration. As most of RNA XIOS graphs are highly connected, an increase in graph size can result in a large increase in the time required to completely sample the fingerprint. Fortunately, the motif sampling is completely parallelizable; any number of processors can independently sample subgraphs from the query, and the time required per query graph is modest. Furthermore, our results suggest that a complete fingerprint may not be necessary; that even incomplete fingerprints (such as fingerprints derived from structures where part of the structure has been removed) are sufficient to identify topologically

similar structures. The question of whether absolutely every subgraph has been detected, which a sampling strategy cannot guarantee, is therefore somewhat moot.

Experimental determination of RNA structures by X-ray crystallography or NMR is difficult, and a relatively small number of complete structures are available. Instead, structures are often predicted using a combination of biochemical information (chemical modification, nuclease sensitivity, and mutational sensitivity), secondary structure prediction, and phylogenetic conservation (covariance). This results in "known" structures that are incomplete (missing important stems) or inaccurate (containing stems that do not exist, or are unimportant in the function of the RNA). It is therefore important that the structural/topological comparison be robust with respect to incompleteness or error in the structures, a salient characteristic of the RNA fingerprint comparison we describe here. The extended-fingerprint Jaccard Similarity correctly identifies topologically similar RNAs across a broad range of sizes, and biological functions, but its potential application is far more general. RNA structure prediction is commonly judged to be 60 to 80 percent accurate (75,78,146). The ability of the RNA fingerprint to correctly identify/classify structural topologies even when 30% or more of the true stems are removed (Figure 3.3F), suggests that this approach can be applied to broadly search for topologically similar structures based on structures predicted from sequence (work in progress).

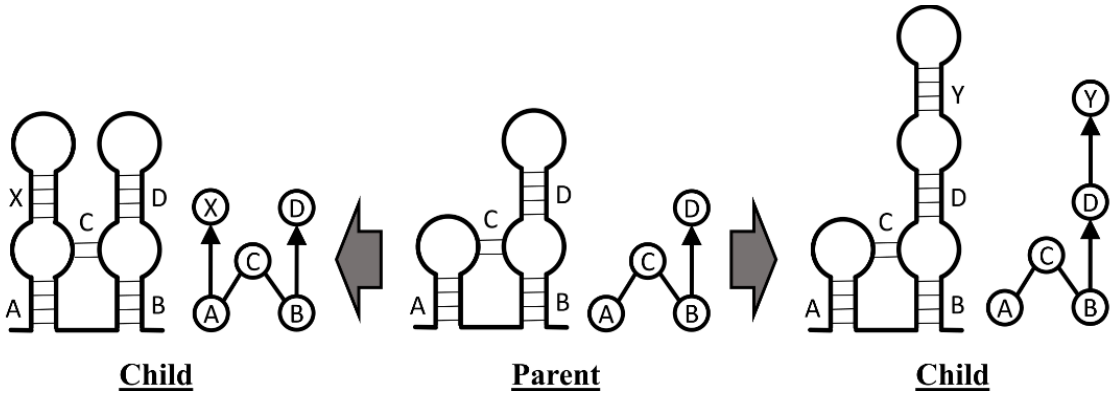


Figure 3.1 Parent-Child relationships. The parent graph is a 4-stem motif; two different child graphs are created by adding one stem to the parent graph.

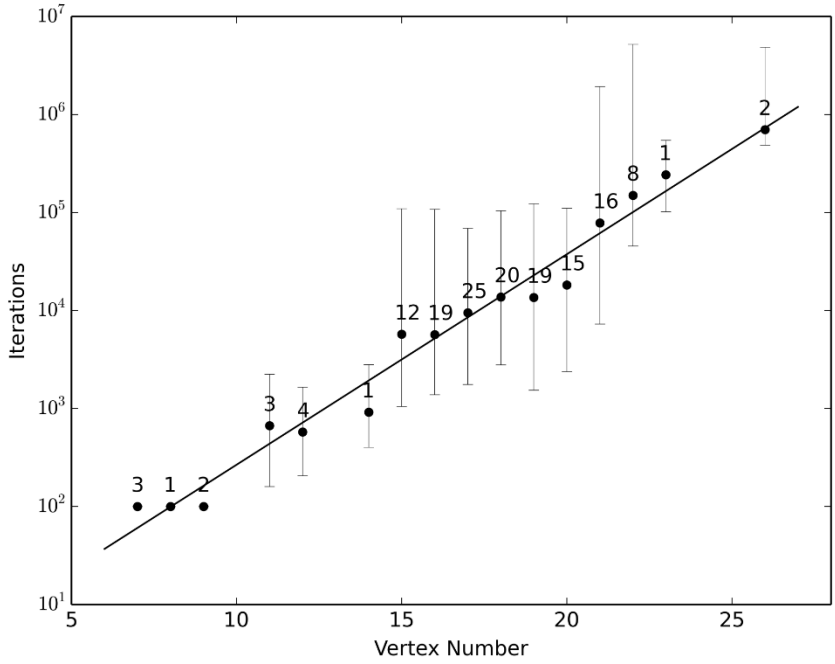


Figure 3.2 Scaling of sampling with graph size. Fingerprints for 151 RNA graphs in the curated set were determined multiple times (10 times per RNA graph) by random sampling. Numbers above the dots indicate the number of different graphs with the same size (vertex number); each dot represents the average iterations needed to determine the complete fingerprint for this specific size group, with bars showing the maximum and minimum iterations as well.

Figure 3.3 Classification performance of similarity functions. Pairwise similarities were calculated, using the indicated similarity functions, for all RNAs in the curated dataset and ranked from high to low. A pair of RNAs from the same curated family is considered a positive match; otherwise they are considered to be a negative match. In all panels, the dashed line indicates the simple fingerprint, and the solid line the extended fingerprint. The AUC for the simple and extended fingerprints, respectively, are indicated in parentheses, below. (A) Intersection Similarity (AUC simple, 0.759; extended, 0.746), (B) Cosine Similarity (0.867; 0.753), (C) Dice Similarity (0.821; 0.864), (D) Hamming Similarity (0.789; 0.834), and (E) Jaccard Similarity (0.870; 0.952). (F) Classification after random removal of vertices from RNA graphs. All RNAs (except for tRNA and 5S rRNA which are too small for 70% stem removal) are included. The five lines show ROC curves with differing fractions of stems removed (AUC in parentheses): (0) no stem removal (AUC=0.909), (1) 10% stem removal (0.844), (2) 30% stem removal (0.810), (3) 50% stem removal (0.691), and (4) 70% stem removal (0.605).

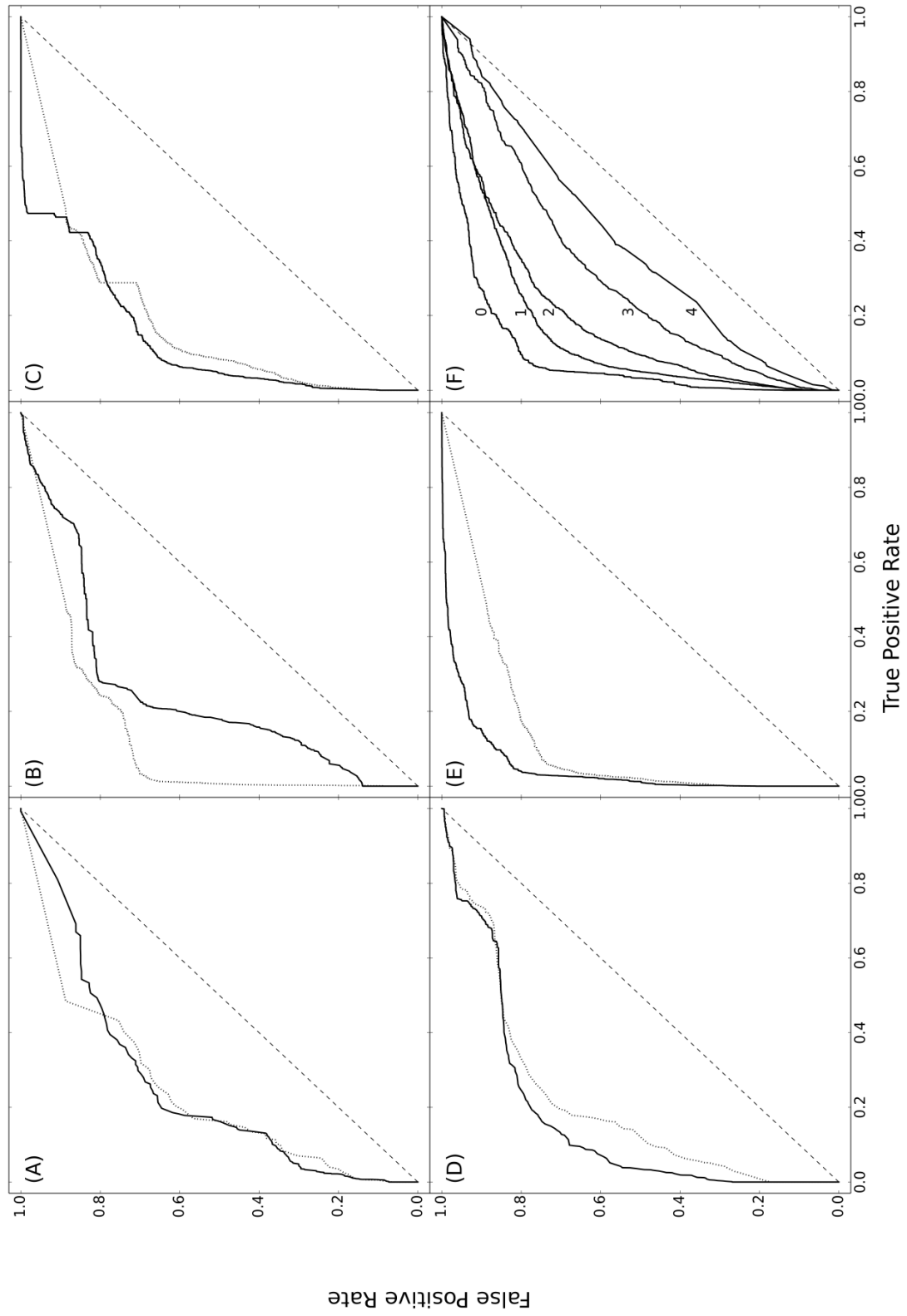

(A)
(B)
(C)
(D)
(E)
(F)
0
1
2
3
4
False Positive Rate
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

Figure 3.3

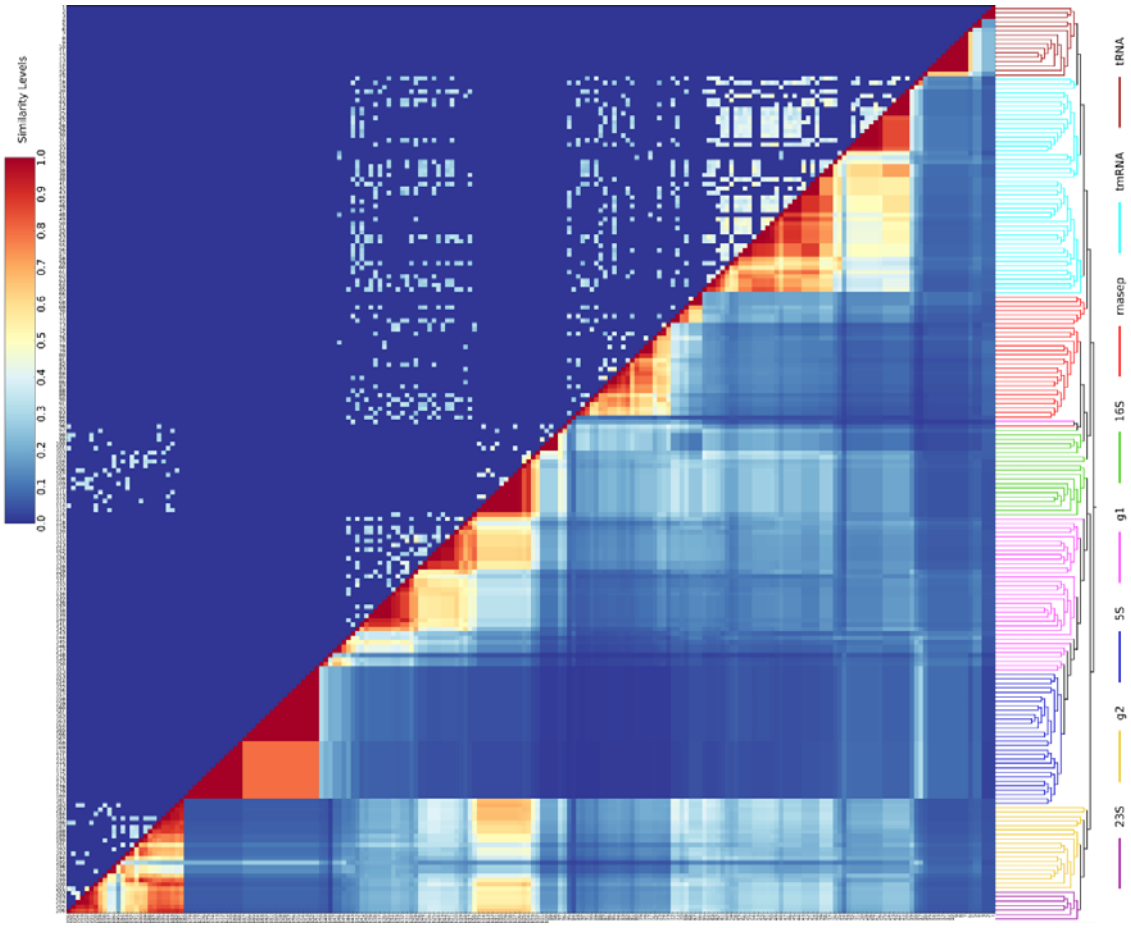


Figure 3.4 Extended-fingerprint Jaccard similarity between biological RNAs. Upper triangle. Sequence identity. Lower triangle. Extended-fingerprint Jaccard Similarity of all the curated RNA structures (see Figure 3.7 and Table 3.7 for IDs). Sequence identity is shown in color, ranging from 0 (blue) to 1 (red) at steps of 0.1. A neighbor-joining dendrogram calculated according to the extended-fingerprint Jaccard similarity is shown on the right side of the heat map.

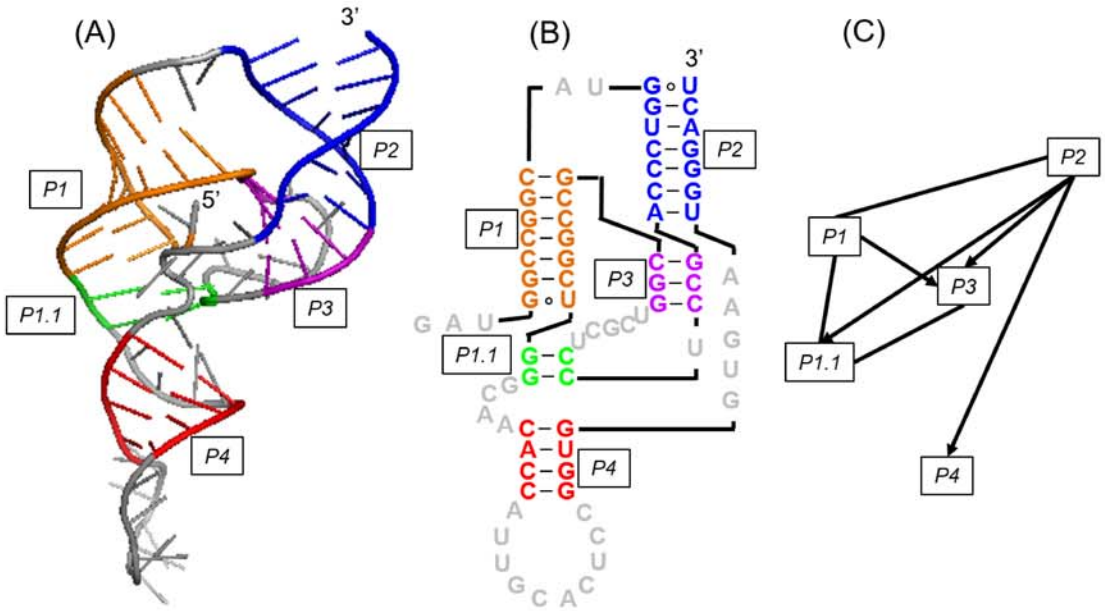


Figure 3.5 XIOS RNA graph representation of a Hepatitis D Virus (HDV) ribozyme RNA. The 3D and 2D structures are adapted from Ke et al (30). (A) 3D structure; (B) 2D structure; (C) XIOS graph representation.

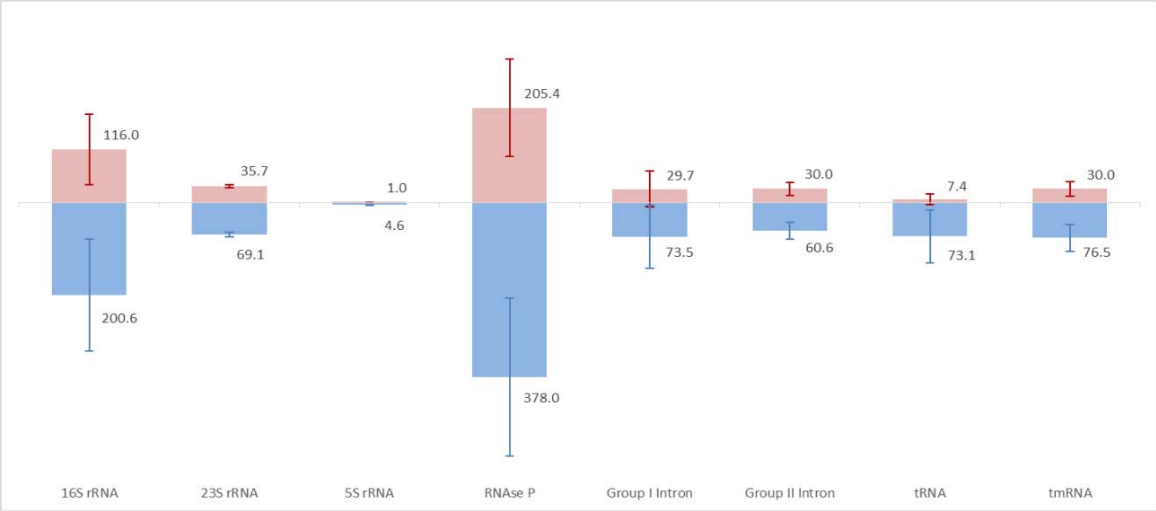


Figure 3.6 Numbers of motifs in Simple and Extended Fingerprints. The average number of motifs in the Simple Fingerprint (upper pink bars, determined by subgraph sampling) or the Extended Fingerprint (lower blue bars, determined by combining sampled subgraphs and all the ancestral subgraphs cataloged in the motif library) in different RNA families a shown beside the corresponding bar. Error bars show the standard deviation of number of motifs.

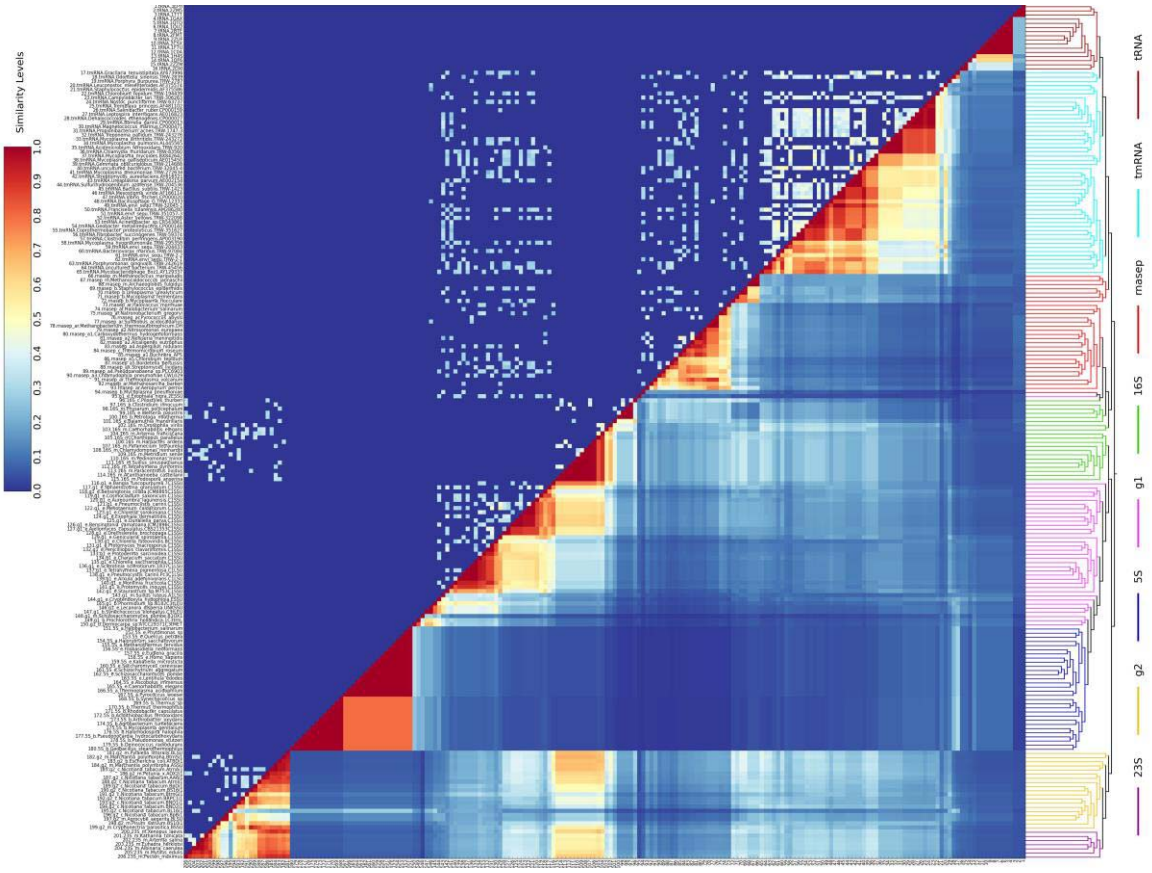


Figure 3.7 Heat map dendrogram. (Same as Figure 3.4, but with RNA names shown) This figure shows the heat map dendrogram of sequence similarity (upper-left triangle) and fingerprint similarity (Extended Jaccard Similarity, lower-right triangle) of all the curated RNA structures (represented by IDs followed by their names, corresponding to Table 3.7). Similarity is shown in different colors, ranging from 0 (blue) to 1 (red) at steps of 0.1. A neighbor-joining tree calculated according to the fingerprint similarity is shown on the right side of the heat map (branching showing the tree topology but branch lengths are not the true distances between nodes).

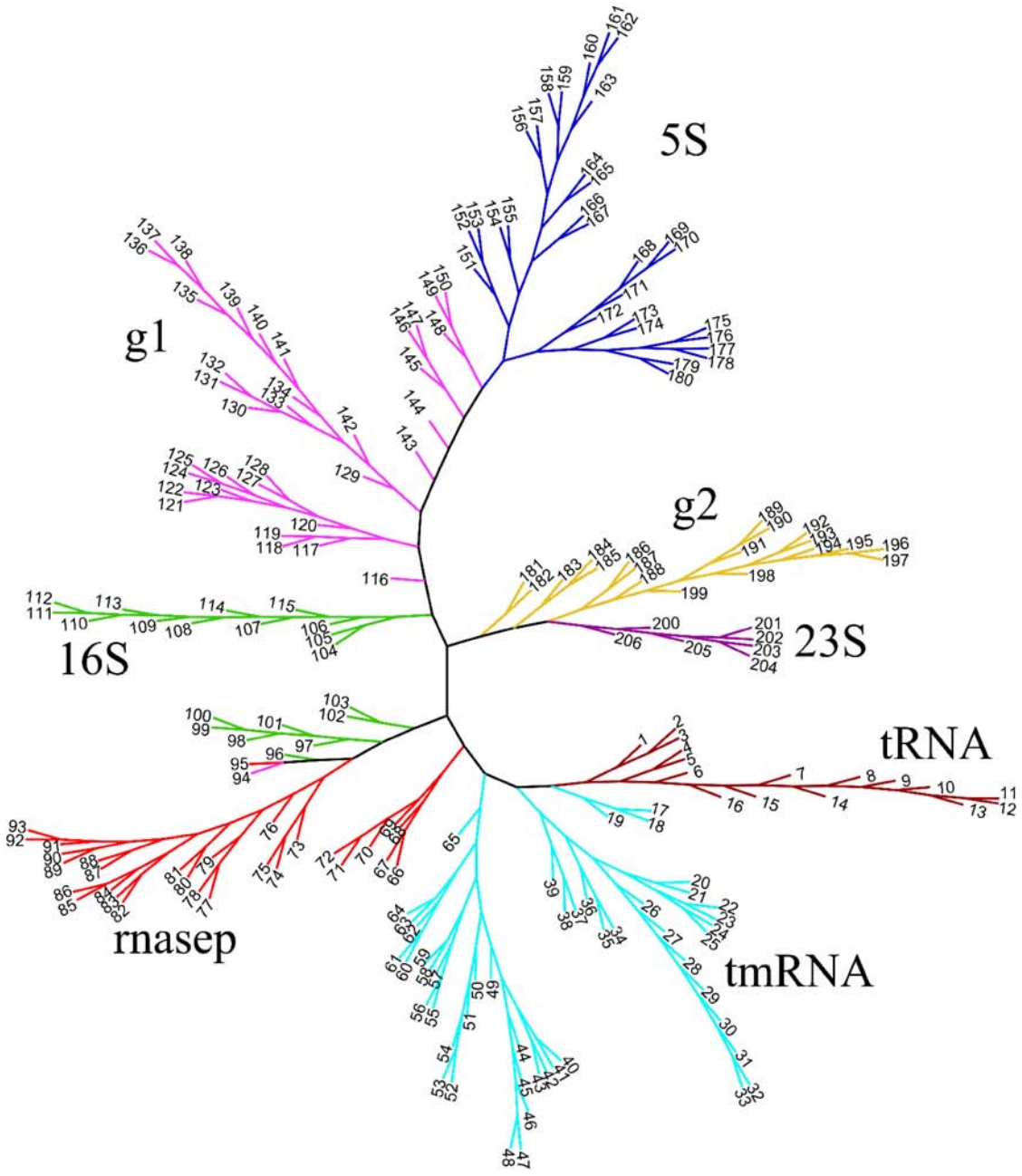


Figure 3.8 Neighbor-joining tree showing the classification using Extended Jaccard Similarity. See Figure 3.7 and Table 3.7 for IDs. The labeled branch is a misclassified RNA: Group I Introns from Exophiala nigra (eukaryotic nucleus) (94).

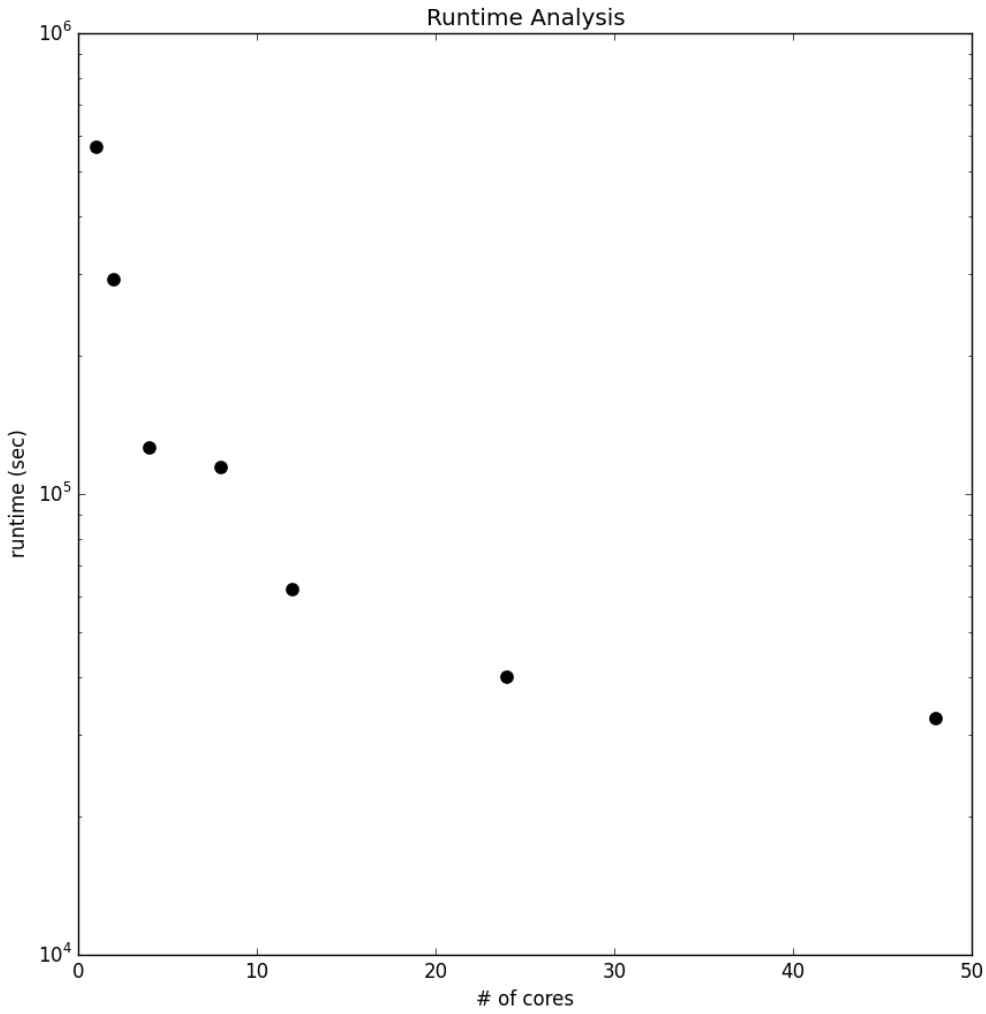


Figure 3.9 Runtime analysis of the subgraph random sampling algorithm.

Table 3.1 Subgraph random sampling pseudocode.

Algorithm: Subgraph Random Sampling

Input: Query graph $G = (V, E)$, subgraph size n
Output: Sampled subgraph $S = (V_s, E_s)$

Select a random vertex $v_i \in V$
Initialize the set of vertices $V_s = \{ v_i \}$
Initialize the set of edges $E_s = \{ \}$

WHILE $| V_s | < n$ DO
  Identify N_{Vs} , the vertices adjacent to V_s
  IF $N_{Vs} == \{ \}$ DO
    BREAK
  ELSE DO
    Select a random vertex v_j from N_{Vs}
    Update $E_s = E_s \cup \{ (v_i, v_j) \} \; \forall v_i, v_j \in V_s$
    Update $V_s = V_s \cup \{ v_j \}$
  END IF
END WHILE
RETURN subgraph $S = (V_s, E_s)$

Table 3.2 RNA fingerprint similarity functions. X and Y are fingerprints of the two structures being compared.

Similarity Function	**Definition**
Intersection	$S_B(X,Y) = \lvert X \cap Y \rvert$
Cosine (147)	$S_C(X,Y) = \frac{\lvert X \cap Y \rvert}{\sqrt{\lvert X \rvert \lvert Y \rvert}}$
Dice (148,149)	$S_D(X,Y) = \frac{2\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert}$
Hamming (150)	$S_H(X,Y) = \lvert (X \cap Y) \cup (X \cup Y)^C \rvert$
Jaccard (151)	$S_J(X,Y) = \frac{\lvert X \cap Y \rvert}{\lvert X \cup Y \rvert}$

Table 3.3 Curated RNA structures. Curated RNA structures with graph characteristics (vertex number, edge number, average degree, and curation description).

RNA Family	N	Vertices	Edges	Average	Description
tRNA	16	7.5(0.9)	15.2(3.9)	4.0(0.5)	Transfer RNAs with resolution < 3 Å from the Protein Data Bank (PDB) (123,152). Base pairing information calculated with RNAView (153). The following PDB IDs are included: 1C0A, 1F7U, 1GAX, 1H4S, 1QF6, 1QTQ, 1QU2, 1TTT, 2BTE, 2CSX, 2DXI, 2FMT, 2ZM5, 2ZUF, 2ZZM, 3EPH.
RNase P RNA	29	16.7(3.1)	72.8(26.0)	8.5(1.7)	Representative Ribonuclease P RNA structures from the classes enumerated by Ellis and Brown (27). Secondary structures and pseudoknots were assigned according to Ellis and Brown (154).

Table 3.3 continued

tmRNA	49	14.0(2.5)	44.3(14.4)	6.2(1.2)	Transfer-messenger (10Sa) RNA. Aligned tmRNA sequences and structural assignments were obtained from Mao et al (155).
Group I Intron RNA	36	17.4(3.6)	26.2(9.4)	2.9(0.7)	Group I Self-Splicing Intron RNA. Sequences and structural assignments were obtained from the Comparative RNA Web site (156); the shortest and longest 10% in length were removed to avoid incomplete or poorly annotated sequences. Containing 3 subgroups: b (bacteria), e (eukaryotic nucleus), and m (eukaryotic mitochondria).

Table 3.3 continued

Group II Intron RNA	19	21.0(4.8)	42.9(17.7)	4.0(0.7)	Group II Self-Splicing Intron RNA. Sequences and structural assignments were from the
5S rRNA	30	4.6(0.5)	5.1(1.0)	2.2(0.2)	5S Ribosomal RNA sequences and structural assignments were obtained from CRW Site (156). Containing 3 subgroups: a (archaea), b (bacteria), e (eukaryotic nucleus).
16S rRNA	20	51.9(14.7)	171.6(74.0)	6.3(1.2)	16S Ribosomal RNA sequences and structural assignments were obtained from CRW Site (156). Containing 4 subgroups: b (bacteria), c (eukaryotic chloroplast), e (eukaryotic nucleus), and m (eukaryotic mitochondria).

Table 3.3 continued

23S rRNA	7	50.4(8.4)	95.0(19.2)	3.7(0.1)	23S Ribosomal RNA sequences and structural assignments were obtained from CRW Site (156). Containing 1 group: m (eukaryotic mitochondria).

Table 3.4 Classification performance of Extended Fingerprint Jaccard Similarity for 8 curated families.

RNA Family	Area Under Curve (AUC)
5S rRNA	1.000
16S rRNA	0.988
23S rRNA	0.989
RNase P	0.964
group I Intron	0.812
group II Intron	0.935
tRNA	1.000
tmRNA	0.985
total	0.961

Table 3.5 Classification performance for expanded graphs using different similarity functions. The value outside of parentheses is area under curve (AUC) for expanded graphs from curated structures; the value inside of parentheses is AUC for decoy graphs, which are considered random and used as controls.

Similarity Function	**SimFP total (decoy only)**	**ExtFP total (decoy only)**
Intersection Similarity	0.699 (0.501)	0.694 (0.481)
Cosine Similarity	0.666 (0.524)	0.654 (0.560)
Dice Similarity	0.806 (0.517)	0.827 (0.500)
Hamming Similarity	0.698 (0.542)	0.721 (0.563)
Jaccard Similarity	0.794 (0.515)	0.840 (0.524)

Table 3.6 Run time analysis. This table shows the average runtime (unit: seconds, rounded to the nearest integer) versus different number of cores for the subgraph sampling algorithm to calculate the fingerprints in each functional RNA family. The runtime reduces as the number of cores increases.

# of cores / Family	1	2	4	8	12	24	48
16S	59289	30113	14916	11563	5598	4267	2976
23S	3638	2533	1088	465	389	192	211
5S	0	0	0	0	0	0	1
g1	3323	1837	880	549	327	232	194
g2	3265	1728	884	509	300	219	196
rnasep	190229	97012	45339	32373	18631	14114	10434
tRNA	1165	540	269	175	101	75	52
tmRNA	3627	1822	967	662	331	221	204

Table 3.7 Complete list of curated RNA structures used in this study. For Group I and II Introns and 5S, 16S, and 23S rRNAs, the subclasses are represented by single letters as follows: a (archaea); b (bacteria); c (cellular components); e (eukaryotic nuclei); m (mitochondria). For RNase P RNA, the subclasses are represented by single or double letters as follows: ar, and m (archaea); a1, a2, a3, a4, a5, ax, b, and c (bacteria). The number of bases in the sequence, the number of vertices, number of edges, and number of pseudoknots are listed under N_base, N_vertex, N_edge, and N_pknot.

ID	Class	Sub-class	Genus	Species	Suffix	N_base	N_vertex	N_edge	N_pknot
1	tRNA	NA	NA	NA	3EPH	68	9	23	9
2	tRNA	NA	NA	NA	2ZM5	75	9	23	9
3	tRNA	NA	NA	NA	1TTT	75	10	23	9
4	tRNA	NA	NA	NA	1GAX	74	7	13	5
5	tRNA	NA	NA	NA	1QTQ	74	7	13	5
6	tRNA	NA	NA	NA	1QU2	74	7	13	5
7	tRNA	NA	NA	NA	2BTE	82	7	13	5
8	tRNA	NA	NA	NA	2FMT	76	7	13	5
9	tRNA	NA	NA	NA	2ZUF	77	7	13	5
10	tRNA	NA	NA	NA	2CSX	74	7	13	5
11	tRNA	NA	NA	NA	1F7U	74	7	13	5
12	tRNA	NA	NA	NA	1C0A	76	7	13	5
13	tRNA	NA	NA	NA	1H4S	76	7	14	5
14	tRNA	NA	NA	NA	1QF6	75	8	17	5
15	tRNA	NA	NA	NA	2ZZM	87	7	13	4
16	tRNA	NA	NA	NA	2DXI	74	7	13	4
17	tmRNA	NA	Gracilaria	tenuis-tipitata	AY673996	396	6	11	1
18	tmRNA	NA	Odontella	sinensis	TRW-2839	371	7	16	1
19	tmRNA	NA	Porphyra	purpurea	TRW-2787	323	8	19	1
20	tmRNA	NA	Leuconos-toc	mesen-teroides	AF375574	307	12	27	5
21	tmRNA	NA	Staphylo-coccus	epider-midis	AF375586	311	11	23	4
22	tmRNA	NA	Chlorobium	tepidum	TRW-194439	404	13	27	4
23	tmRNA	NA	Campylo-bacter	lari	TRW-306263	359	13	35	4

Table 3.7 continued

24	tmRNA	NA	Nostoc	punctiforme	TRW-63737	390	15	42	5
25	tmRNA	NA	Tremblaya	princeps	AF481102	264	11	28	3
26	tmRNA	NA	Salinibacter	ruber	CP000159	369	13	35	4
27	tmRNA	NA	Leptospira	interrogans	AE016823	350	13	35	4
28	tmRNA	NA	Dehalococcoides	ethenogenes	CP000027	352	13	35	4
29	tmRNA	NA	Borrelia	garinii	CP000013	363	14	38	4
30	tmRNA	NA	Magnetococcus	marinus	CP000471	366	12	30	2
31	tmRNA	NA	Propionibacterium	acnes	TRW-1747-3	353	13	42	3
32	tmRNA	NA	Treponema	pallidum	TRW-243276	354	14	47	4
33	tmRNA	NA	Mycoplasma	arthritidis	TRW-243272	394	14	47	4
34	tmRNA	NA	Mycoplasma	pulmonis	AL445565	387	14	47	4
35	tmRNA	NA	Acidimicrobium	ferrooxidans	TRW-920	341	14	47	4
36	tmRNA	NA	Chlamydia	muridarum	TRW-83560	421	14	47	4
37	tmRNA	NA	Mycoplasma	mycoides	BX842642	411	15	52	4
38	tmRNA	NA	Mycoplasma	gallisepticum	AE015450	408	16	56	4
39	tmRNA	NA	Gemmata	obscuriglobus	TRW-214688	412	16	56	4
40	tmRNA	NA	uncultured	bacterium	TRW-32045-4	369	16	67	5
41	tmRNA	NA	Mycoplasma	pneumoniae	TRW-272634	387	14	40	5
42	tmRNA	NA	Streptomyces	aureofaciens	AY616521	382	14	40	5
43	tmRNA	NA	Ureaplasma	parvum	AE002154	413	14	40	5

Table 3.7 continued

44	tmRNA	NA	Sulfurihydrogenibium	azorense	TRW-204536	351	14	40	5
45	tmRNA	NA	Bacillus	subtilis	TRW-1423	363	15	53	5
46	tmRNA	NA	Mesostigma	viride	AF166114	359	15	53	5
47	tmRNA	NA	Vibrio	fischeri	CP000020	367	15	53	5
48	tmRNA	NA	Bacillusphage	G	TRW-12333	312	11	35	3
49	tmRNA	NA	envi	sequ	TRW-32045-1	355	15	53	5
50	tmRNA	NA	Francisella	tularensis	AM286280	421	16	58	5
51	tmRNA	NA	envi	sequ	TRW-351057-3	355	16	58	5
52	tmRNA	NA	Aster	yellows	TRW-322098	426	16	59	6
53	tmRNA	NA	Acinetobacter	sp	CR543861	360	14	48	3
54	tmRNA	NA	Geobacter	metallireducens	CP000148	356	17	74	6
55	tmRNA	NA	Coprothermobacter	proteolyticus	TRW-351627	353	15	45	6
56	tmRNA	NA	Fibrobacter	succinogenes	TRW-59374	361	15	45	6
57	tmRNA	NA	Clostridium	perfringens	AP003190	358	15	45	6
58	tmRNA	NA	Mycoplasma	hyopneumoniae	TRW-295358	424	16	42	9
59	tmRNA	NA	envi	sequ	TRW-204433	356	14	32	6
60	tmRNA	NA	Bacteriovorax	marinus	TRW-97084	385	16	51	7
61	tmRNA	NA	envi	sequ	TRW-2-2	356	14	33	6

Table 3.7 continued

62	tmRNA	NA	envi	sequ	TRW-2-1	365	15	46	6
63	tmRNA	NA	Porphyromonas	gingivalis	TRW-242619	407	18	71	7
64	tmRNA	NA	uncultured	bacterium	TRW-45456	408	19	77	7
65	tmRNA	NA	Mycobacteriophage	Bxz1	AY129337	437	18	71	4
66	rnasep	m	Methanococcus	maripaludis	NA	233	10	30	1
67	rnasep	m	Methanocaldococcus	jannaschii	NA	252	10	30	1
68	rnasep	m	Archaeoglobus	fulgidus	NA	229	10	30	1
69	rnasep	b	Staphylococcus	epidermidis	NA	401	19	76	1
70	rnasep	b	Ureaplasma	urealyticum	NA	370	19	76	1
71	rnasep	b	Mycoplasma	fermentans	NA	302	15	42	1
72	rnasep	b	Mycoplasma	floccularе	NA	412	15	46	1
73	rnasep	ar	Halococcus	morrhuae	NA	475	23	146	3
74	rnasep	ar	Halobacterium	salinarum	NA	375	19	106	3
75	rnasep	ar	Natronobacterium	gregoryi	NA	474	22	139	3
76	rnasep	ar	Pyrococcus	abyssi	NA	330	16	78	4
77	rnasep	ar	Sulfolobus	acidocaldarius	NA	315	15	68	4
78	rnasep	ar	Methanobacterium	thermoautotrophicum	DH	293	15	68	4
79	rnasep	a2	Nitrosomonas	europaea	NA	285	14	54	4

Table 3.7 continued

80	rnasep	a1	Carboxydothermus	hydrogenoformans	NA	331	16	68	4
81	rnasep	a2	Neisseria	meningitidis	NA	360	16	68	4
82	rnasep	a2	Alcaligenes	eutrophus	NA	341	15	56	4
83	rnasep	a4	Aspergillus	nidulans	NA	385	17	70	4
84	rnasep	c	Thermomicrobium	roseum	NA	350	18	78	5
85	rnasep	a1	Buchnera	APS	NA	376	17	74	5
86	rnasep	a5	Chlorobium	tepidum	NA	381	18	79	5
87	rnasep	a5	Bordetella	pertussis	NA	414	20	92	5
88	rnasep	ax	Streptomyces	lividans	NA	405	18	75	4
89	rnasep	a4	Pseudoanabaena	sp	PCC6903	450	18	74	4
90	rnasep	a3	Chlamydophila	pneumoniae	CWL029	406	19	81	5
91	rnasep	ar	Thermoplasma	volcanum	NA	305	16	70	4
92	rnasep	ar	Methanosarcina	barkeri	NA	371	18	76	4
93	rnasep	ar	Aeropyrum	pernix	NA	330	15	73	3
94	rnasep	b	Mycoplasma	pneumoniae	NA	369	20	89	7
95	g1	e	Exophiala	nigra	2ESSU	445	19	28	5
96	16S	c	Pilostyles	thurberi	NA	1464	63	188	4
97	16S	b	Clostridium	innocuum	NA	1529	72	279	5
98	16S	m	Physarum	polycephalum	NA	1844	65	248	5

Table 3.7 continued

99	16S	e	Weiseria	palustris	NA	1373	64	257	5
100	16S	b	Petrotoga	miotherma	NA	1326	67	273	5
101	16S	e	Balamuthia	mandrillaris	NA	1972	69	262	4
102	16S	m	Drosophila	virilis	NA	783	34	85	2
103	16S	m	Caenorhabditis	elegans	NA	678	32	83	3
104	16S	m	Artemia	franciscana	NA	711	31	74	1
105	16S	m	Chorthippus	parallelus	NA	789	28	55	1
106	16S	m	Harpactes	ardens	NA	950	34	83	1
107	16S	m	Paramecium	tetraurelia	NA	1609	56	194	1
108	16S	m	Chlamydomonas	reinhardtii	NA	1188	44	128	1
109	16S	m	Metridium	senile	NA	1088	45	128	1
110	16S	m	Pedinomonas	minor	NA	1170	54	170	1
111	16S	m	Suillus	sinuspaulianus	NA	1976	69	246	1
112	16S	m	Tetrahymena	pyriformis	NA	1632	53	164	1
113	16S	m	Paracentrotus	lividus	NA	877	35	91	1
114	16S	m	Acanthamoeba	castellanii	NA	1523	61	210	1
115	16S	m	Podospora	anserina	NA	1759	62	214	1
116	g1	e	Bangia	fuscopurpurea	7C1SSU	475	24	49	1
117	g1	e	Sphaerozosma	granulatum	C1SSU	432	18	34	1

Table 3.7 continued

118	g1	e	Bensington-ia	ciliata	JCM6865C1SSU	334	14	20	1
119	g1	e	Cosmo-cladium	saxoni-cum	C1SSU	443	16	27	1
120	g1	e	Aureoum-bra	lagunen-sis	C1SSU	438	21	42	1
121	g1	e	Pneumo-cystis	carinii	C1SSU	404	19	31	1
122	g1	e	Mesotaeni-um	caldario-rum	C1SSU	414	21	33	1
123	g1	e	Chlorella	sorokin-iana	C1SSU	478	21	34	1
124	g1	e	Exophiala	derma-titidis	C1SSU	425	21	34	1
125	g1	e	Dunaliella	parva	C1SSU	394	21	36	1
126	g1	e	Bensington-ia	yamato-ana	JCM2896C1SSU	468	24	39	1
127	g1	e	Ajellomyces	capsula-tus	CBS21353C1SSU	407	19	27	1
128	g1	e	Drechslerel-la	brochop aga	C1SSU	405	16	25	1
129	g1	e	Genicularia	spiro-taenia	C1SSU	382	17	31	1
130	g1	e	Chlorella	luteovi-ridis	BC1SSU	439	21	28	1
131	g1	e	Protomyces	macro-sporus	C1SSU	394	18	25	1
132	g1	e	Penicilliop-sis	clavar-iiformis	C1SSU	388	15	22	1
133	g1	e	Protoderma	sarcinoi-dea	C1SSU	457	19	31	1
134	g1	e	Characium	sac-catum	C1SSU	461	21	35	1
135	g1	e	Chlorella	sac-char-ophila	C1SSU	394	19	30	1
136	g1	e	Sclerotinia	sclerotio-rum	1837C1LSU	320	14	25	1

Table 3.7 continued

137	g1	e	Tetrahymena	pigmentosa	C1LSU	422	19	30	1
138	g1	e	Pneumocystis	carinii	Pc3C1LSU	375	15	25	1
139	g1	e	Arxula	adeninivorans	C1LSU	425	18	29	1
140	g1	e	Monilinia	fructicola	C1SSU	432	17	22	1
141	g1	e	Protomyces	inouyei	C1SSU	353	20	31	1
142	g1	e	Staurastrum	sp	M753C1SSU	410	16	20	1
143	g1	m	Suillus	luteus	A1LSU	356	14	15	1
144	g1	e	Cryptendoxyla	hypophloia	ESSU	448	19	19	1
145	g1	b	Phormidium	sp	N182C3tLEU	269	13	16	1
146	g1	e	Lecanora	dispersa	UNKSSU	311	10	11	1
147	g1	b	Synechococcus	elongatus	C3tLEU	252	13	13	1
148	g1	m	Schizosaccharomyces	pombe	B1OX1	270	8	4	1
149	g1	b	Prochlorothrix	hollandica	1C3trnL	281	13	12	1
150	g1	b	Dermocarpa	sp	ATCC29371 C3tMET	266	13	11	1
151	5S	a	Halobacterium	salinarum	NA	120	5	6	0
152	5S	e	Phytomonas	sp	NA	123	5	6	0
153	5S	e	Quercus	petraea	NA	119	5	6	0
154	5S	a	Halorubrum	saccharovorum	NA	122	5	6	0
155	5S	a	Methanothermus	fervidus	NA	123	5	6	0
156	5S	e	Filobasidiella	neoformans	NA	117	5	6	0

Table 3.7 continued

157	5S	e	Euglena	gracilis	NA	120	5	6	0
158	5S	e	Homo	sapiens	NA	118	5	6	0
159	5S	e	Kabatiella	mi-crosticta	NA	119	5	6	0
160	5S	e	Saccharo-myces	cere-visiae	NA	117	5	6	0
161	5S	e	Schizo-chytrium	aggrega-tum	NA	118	5	6	0
162	5S	e	Schizosac-charomyces	pombe	NA	118	5	6	0
163	5S	e	Lentinula	edodes	NA	119	5	6	0
164	5S	e	Ascobolus	immer-sus	NA	118	5	6	0
165	5S	e	Caenorhab-ditis	elegans	NA	118	5	6	0
166	5S	a	Thermo-plasma	acidophi-lum	NA	122	5	6	0
167	5S	a	Pyrococcus	woesei	NA	123	5	6	0
168	5S	b	Synecho-coccus	sp	NA	119	4	4	0
169	5S	b	Thermus	sp	NA	119	4	4	0
170	5S	b	Thermus	ther-mophilus	NA	120	4	4	0
171	5S	b	Rhodobac-ter	capsula-tus	NA	118	4	4	0
172	5S	b	Acidithio-bacillus	ferrooxi-dans	NA	119	4	4	0
173	5S	b	Arthrobac-ter	oxydans	NA	120	4	4	0
174	5S	b	Agrobacte-rium	tumefa-ciens	NA	119	4	4	0
175	5S	b	Mycoplas-ma	genitali-um	NA	117	4	4	0
176	5S	b	Halor-hodospira	halophila	NA	120	4	4	0

Table 3.7 continued

177	5S	b	Pseudono-cardia	hydro-carbon-oxydans	NA	119	4	4	0
178	5S	b	Pseudomo-nas	stutzeri	NA	117	4	4	0
179	5S	b	Deinococ-cus	radi-odurans	NA	123	4	4	0
180	5S	b	Geobacillus	stea-rother-mophilus	NA	118	4	4	0
181	g2	m	Pylaiella	littoralis	BLSU	2411	28	67	0
182	g2	m	Marchantia	poly-morpha	BtrnSi1	992	36	99	0
183	g2	b	Escherichia	coli	ATBDi1	1894	24	52	0
184	g2	m	Marchantia	poly-morpha	ASSU	1610	24	62	0
185	g2	c	Nicotiana	tabacum	AtrnAi1	712	23	53	0
186	g2	m	Petunia	x	AOX2i1	1356	21	45	0
187	g2	c	Nicotiana	tabacum	AA6i1	698	19	37	0
188	g2	c	Nicotiana	tabacum	AtrnIi1	710	19	36	0
189	g2	c	Nicotiana	tabacum	BpDi1	742	19	37	0
190	g2	c	Nicotiana	tabacum	BS16i1	861	18	35	0
191	g2	c	Nicotiana	tabacum	BtrnGi1	691	18	26	0
192	g2	c	Nicotiana	tabacum	BRPC1i1	739	20	45	0
193	g2	c	Nicotiana	tabacum	BND1i1	1148	17	28	0
194	g2	c	Nicotiana	tabacum	BND2i1	679	19	33	0
195	g2	c	Nicotiana	tabacum	BL16i1	1019	14	21	0

Table 3.7 continued

196	g2	c	Nicotiana	tabacum	BpBi1	753	16	28	0
197	g2	m	Agrocybe	aegerita	BLSU	1762	20	29	0
198	g2	m	Pisum	sativum	BS10i1	934	24	40	0
199	g2	m	Cryphonectria	parasitica	BSSU	2069	20	42	0
200	23S	m	Xenopus	laevis	NA	1634	69	138	0
201	23S	m	Katharina	tunicata	NA	1272	47	85	0
202	23S	m	Artemia	salina	NA	1135	46	84	0
203	23S	m	Euhadra	herklotsi	NA	1021	43	79	0
204	23S	m	Albinaria	caerulea	NA	1031	43	80	0
205	23S	m	Mytilus	edulis	NA	1241	53	99	0
206	23S	m	Pecten	maximus	NA	1401	52	100	0

CHAPTER 4. IDENTIFICATION OF RNA STRUCTURAL ENSEMBLES WITH PSEUDOKNOTS USING COMBINATION OF MULTIPLE PREDICTION PROGRAMS

4.1 Introduction

Cellular RNAs, both coding and non-coding, adopt complex folded structures in vivo. The structure of RNA is usually conserved along with its function, however RNA structures are difficult to determine by traditional experimental approaches, such as NMR or X-ray crystallography. An important alternative method for examining RNA structure is computational prediction.

The most commonly used RNA structure prediction approaches use dynamic programming (DP) (56-61), and incorporate experimentally determined nearest-neighbor energy parameters (54). There are several DP-based program suites that predict both minimum free energy and near minimum free energy RNA structures, including Mfold/UNAfold (3,62,63), RNAstructure (64-66), and ViennaRNA (67-69). In nature, the folding of RNA is spontaneous because base-pairing and stacking reduce the free energy of RNA molecules. However, RNA molecules are dynamic and instead of adopting a single folded conformation, they form an ensemble of interconverting structures with near-minimum free energies. McCaskill developed partition function algorithm (McCaskill 1990) that samples the ensemble of RNA structures from Boltzmann distribution and calculates the probability of each structure within the ensemble from its free energy.

Mfold generates a thermodynamically optimal base-paired structure, using dynamic programming; this approach can be extended, using multiple tracebacks in the dynamic programming matrix, to produce suboptimal structures as well. UNAFold (63), an extension of mfold, also includes partition function calculations which permit, among other things, the determination of base-pairing probabilities for each base. The Fold program, like Mfold, uses dynamic programming to predict minimum free-energy structures (36). Other programs that incorporate partition function to calculate base-pairing probability include: ProbablePair, which predicts a structure that incorporates the base-pairs whose probability exceeds a threshold (157); AllSub, which computes all the possible suboptimal structures plus the optimal structure (158); MaxExpect, which shows only the RNA structures with the highest base-pairing probabilities (146); stochastic, which samples RNA structures from the Boltzmann ensemble according to their probability of occurrence (159); and ProbKnot, which predicts a maximum expected accuracy structure using base-pairing probabilities calculated by the partition function algorithm (78), etc.

The ViennaRNA program RNAfold uses dynamic programming to compute base-pairing probabilities using the McCaskill partition function algorithm and produces both the MFE structure and suboptimal structures. Another program, RNALfold, computes the locally stable structure within a region of an RNA sequence. Other approaches also use partition function to calculate base-pairing probabilities, for example, Sfold (70,71) calculates a centroid structure from Boltzmann ensemble, which has the smallest base-pair distance to the other structures in the ensemble.

Originally, dynamic programming-based program suites, such as ViennaRNA and RNAstructure, did not include pseudoknot prediction. This is because the prediction of minimum free-energy structures containing pseudoknots requires calculation of non-nested base-pairs which significantly increases the complexity of the calculation (76). In general, due to memory and time limitations, dynamic programming based programs do not predict pseudoknots; however, some program suites have extended secondary-structure prediction to include pseudoknots by modifications incorporating various heuristics into their algorithms. ViennaRNA includes the program RNAPKplex (77), which decomposes a secondary structure into two parts and separately calculates the minimum free energy of each part. The two parts include a pseudoknot-free structure that includes accessible (unpaired) bases, and an additional stem formed within the accessible region to form a pseudoknot with a stem in the pseudoknot-free structure. The calculation of pseudoknot energy is recursive and with complexity O(n6); when the length of the accessible region limited to w, the computational time for RNAPKplex is O(n3 + n2w4). In the RNAstructure program suite, ProbKnot predicts pseudoknot-free structures using base-pairing probabilities calculated by the partition function algorithm, and recursively searches for base-pairs with the highest pairing probabilities to yield a maximum expected accuracy structure that contains pseudoknots in O(n3) time (78).

Other approaches for solving the pseudoknot prediction problem have been developed recently. However, these programs often predict only certain types of pseudoknots. The best-characterized type is the H-type pseudoknot, sometimes called simple pseudoknot, which is the interaction between a hairpin loop and the region outside the loop. Anoth-

er widely studied pseudoknot is the kissing hairpin, which is the interaction between the loop regions of two hairpins loops. DotKnot predicts H-type pseudoknots and kissing hairpins by extracting high probability paired regions from an initial near minimum free-energy structure prediction, assembling a list of candidate pseudoknots, and evaluating the pseudoknot loop entropies using parameters developed by Cao and Chen(79,80). DotKnot has complexity O(n3), with the loop length of the pseudoknots limited. Another package, pknotsRG (130), originally predicted H-type pseudoknots and has been superseded by the novel program, pKiss (81,82), that includes kissing hairpins using a heuristic strategy with complexity O(n5).

In addition to for dynamic programming and partition function based approaches, other RNA structure prediction approaches include grammar-based methods, such as CONTRAfold (72,73). CONTRAfold computes base-pairing probabilities using conditional log-linear models (CLLM), a generalization of stochastic context-free grammars (SCFG), using free-energy scoring of RNA structural features (stems, loops, bulges, etc). The CONTRAfold model is used in other RNA structure prediction packages such as CentroidFold (74) and IPknot (75). These programs compute base-pairing probabilities using the McCaskill partition function algorithm, CONTRAfold grammar-based methods, or other approaches, and then predict a structure with maximizeald base-pairing probability using integer programming (IP) (IPknot), or by a generalized centroid approach (CentroidFold).

RNA molecules are ensembles of interconverting structures with different topologies, presumably near the minimum free-energy. RNA structures predicted by different programs are each only partially correct due to the incompleteness of the nearest neighbor

energy model, imprecision in energy parameters, interactions with proteins and ions in the cell, and temperature and ionic strength effects. Since individual program each have limitations, in this work we have investigate whether combining the results of multiple programs improves the accuracy of predicted structures. The combined structure compresses several alternative structures into one: it removes the redundant base-pairing regions that exist in more than one alternative structures, and retains all the base-pairing information. We have tested 24 state-of-the-art RNA structure prediction programs on a gold-standard set of functional RNA sequences, most of which include biologically validated pseudoknots,using 327,679 combinations of RNA structure prediction programs. . This comprehensive comparison confirms previous findings that predicted structures are highly sensitive to the prediction methods, and allows the identification of program combinations that give the most accurate structures. The best combinations vary with different trades-offs between prediction recall and precision.

4.2 Materials and Methods

4.2.1 Curated RNA families

A set of curated RNA structures have been collected from the literature and a variety of biological databases (106) and is extended in this work (Table 3.3). This set includes 206 structures of transfer RNA, Ribonuclease P RNA, transfer-messenger RNA, group I and group II self-splicing introns, and 5S, 16S and 23S ribosomal RNA. The structures in this curated set have been reviewed to ensure that the reported structures are as accurate as possible given existing experimental data, and incorporating expert opinion. The cu-

rated structures have been screened to ensure that no pair of structures has greater than 50% sequence identity.

4.2.2 RNA structure prediction by different programs

A list of 24 RNA structure prediction programs (Table 4.1, UNAFold with 9 different parameter combinations, plus 15 other programs) have been tested on the set of curated RNASome of the programs predict ensembles of structures or allow sets of near minimum free-energy structures (Table 4.1). For example, UNAFold predicts both the minimum free energy (MFE) and near minimum free-energy structures. The set of near minimum free-energy varies with the limitation on ΔΔG (energy difference with respect to the MFE structure) and W (window size; each suboptimal predicted structure has at least W pairs of bases that are different from all other structures; moreover, each of the W pairs must have at least W bases in its position away from any pairs in other structures) (62).

4.2.3 Predicted structure evaluation: precision, recall, and F1 score

A predicted structure can be evaluated by evaluating the precision, recall, and accuracy (F1 score) of stems using the corresponding curated structure as the gold standard.

Precision is the proportion of stems in a predicted structure that match curated stems.

$$Precision = \frac{TP}{TP + FP}$$

where, the True Positive stems, TP, are defined as the number of predicted stems that match at least one curated stem. Stems are considered to match when each half of one stem shares at least one base that is the same as the corresponding half of the other stem. FP, False Positives, are defined as the number of predicted stems that do not match any curated stem. The sum of TP and FP is the total number of predicted stems.

Recall calculates the fraction of curated stems that match predicted stems. FN, False Negatives, are defined as the number of curated stems that do not match any predicted stem. In recall calculation, the sum of TP and FN is the total number of curated stems.

$$Recall = \frac{TP}{TP + FN}$$

The F1 score measures the accuracy of a predicted structure giving equal weight to precision and recall.

$$F1 = \frac{2 * precision * recall}{precision + recall}$$

4.2.4 Correlation analysis of structure prediction programs

Structure prediction performance with using the 24 conditions and programs described above shows different patterns across the 8 RNA families (Table 4.4), which conserve different structural characteristics. It Specific programs/conditions has presumably work better with structures with specific characteristics, and perform better in the families with such structural characteristics. However, RNA molecules are also complex molecules with multiple structural patterns, which, might not be perfectly captured by a sin-

gle program. Predictions based on multiple independent programs have the potential to provide prediction that better match the individual characteristics of specific families, and thus increase recall, but may also generate noise (false positives, over-prediction) and reduce precision. The selection of a reasonable number of independent programs is like making a pot of stew: if the programs are seasonings, they need to be diverse enough for a decent flavor (recall); they also need to be limited to avoid over spicing (precision).

The program performance (precision, recall, F1 score) is calculated from the proportion of matching stems; the independence of the programs can be analyzed by the correlations in matching to the curated stems. There are in total 4,455 stems in the 206 curated RNA sequences. We generated a 24 (programs) x 4,455 (stems) data matrix; each row is a binary vector of 4,455 elements with one indicating that the corresponding curated stem matched by at least one stem predicted by the program (row) and zero indicating that the corresponding curated stem cannot be matched to any stem predicted by that program. Correlation between any two programs were calculated, and used to compute the distance which is the inverse of correlation. A UPGMA (Unweighted Pair Group Method with Arithmetic Mean) tree was generated from the correlation. In the UPGMA tree, the branches started from single programs; branches with least distances were joined to create a new branch, and a new distance was calculated; this process continued until all the branches were joined to a common ancestor (root of the tree).

4.2.5 Combination of alternative structures

The alternative foldings predicted for the same RNA sequence often share a large proportion of identically base-paired regions (Figure 4.1). We remove this redundancy to generate a combined structure by combining overlapping stems in the predicted structures into single stems. Two stems are considered to overlap when both of their half-stems completely or partially overlap in the sequence and their centers (the mean of the position of the first and last paired bases) differ by no more than 3 bases. The union of the sequence positions of both the paired bases is used as the coordinates of the combined stem. Structure combination is complete when no overlapping can be found between any two stems in the predicted structures. Figure 4.1 shows an example of structure combination. The MFE structure contains two base-paired regions: 3-7:22-26, and 11-12:17-18. The near-MFE structure also has two base-paired regions: 2-5:23-26, and 7-9:16-18. The first stem in each predicted structure overlap on both the left and right half stems, bases 3-7 vs 2-5 and 22-26 vs 23-26, respectively. The second base-paired regions overlap on only the right half stem, bases 17-18 vs 16-18, while the left half stems, bases 11-12 vs 7-9, do not overlap. Moreover, the left half stem of second base-paired region in 1, right (bases 7-9) overlaps with the left half stem of 1, left (bases 3-7) by one base.

Structure combination was performed on all possible sets of alternative structures generated of the 24 program/conditions considered here. The 24 programs/conditions have, in total, 327,679 combinations; for each RNA sequence, 327,679 combined predicted structures have been generated.

4.3 Results

4.3.1 Average precision, recall, and F1 score of 24 RNA structure prediction programs

Table 4.4 shows the average performance (precision, recall, and F1 score) over all the RNA sequences, and for each structural class, for different RNA structure prediction programs/conditions. For a single program, the best average precision, recall, and F1 score in the 206 RNA sequences are 0.763±0.246 (mean±standard deviation, Probable-Pair in RNAstructure), 0.769±0.193 (sfold) and 0.577±0.241 (ProbKnot in RNAstructure), respectively.

A precision of 0.763±0.246 means that more than 75% of the gold-standard curated stems are correctly predicted; recall of 0.769±0.193 indicates that over 75% of the predicted stems correspond to known stems in the curated biological structures. A predicted structure that comprises 75% of the correct stems is sufficiently complete for classification using topological methods (Huang and Gribskov, in preparation). We have found that in structures with 70% of correct stems we are able to discriminate between different RNA functional families according to their structures (AUC > 0.8; data not shown). Using a single program, we are confident to obtain 75% of correct stems in structure prediction; however, if we can find an optimal combination of results from multiple programs, we might be able to push the performance to a higher level.

4.3.2 Correlation analysis

To further analyze the relationship between programs, a UPGMA tree was calculated from the correlation between different RNA structure prediction programs/conditions (Figure 4.2). The UNAFold programs with different parameter settings, the programs

from RNAstructure package, and other programs (grammar-based programs and statistical sampling-based programs), group together in major branches of the tree. Programs on different branches are more independent from each other, and when structures predicted by these programs are combined, we expect they should show better prediction performance.

4.3.3 F1 score of the 20 best performing program combinations

Table 2 shows the top 20 best performing combinations based on average F1 scores across the curated families. For each combination, the average ranking is calculated by summing the rank of the program within each family, and dividing by the number of families. Predicted RNA structures from 327,679 program combinations were considered. Program combination CONTRAfold in IPknot combined with ProbKnot in RNAstructure (n.p), ProbKnot in RNAstructure (p), CentroidFold combined with ProbKnot in RNAstructure (c.p), McCaskill in IPknot combined with ProbKnot in RNAstructure (i.p) are the top 4 in average F1 ranking, with n, c, and i in one group of the UPGMA tree, and p in another group (Figure 4.2), which agrees with our expectation that combination of more independent programs should have better prediction performance. More importantly, n, c, and i are pseudoknot prediction programs.

4.3.4 Pareto Frontier: potential optimal solutions for precision versus recall

Figure 4.3 shows the precision and recall of the program combinations. For each family, a Pareto frontier is shown. Each point on the Pareto frontier exceeds all the points with

smaller vertical coordinates in precision, and exceeds all the points with larger vertical coordinates in recall. The full list of points on the Pareto frontier can be found in the Tables 4.7 – 4.14. The lists of program combinations with the top precision and recall can also be found in Table 4.5 and Table 4.6.

Table 4.3 shows the frequency of programs on the Pareto Frontier, as a reference to choose programs depending on the need of users – higher F1, precision, or recall. Since F1 score is skewed towards precision, similar conclusions can be drawn for both of these two statistics. If one's goal is to obtain predicted structures with higher precision or F1 score, programs such as CONTRAfold in IPknot (n), CentroidFold (c), McCaskill in IPknot (i), ProbablePair (r), AllSub in RNAstructure (a), pKiss (k), and ProbKnot in RNAstructure (p) are better options. To obtainpredicted structures with higher recalls, i.e., with lower numbers of false positives, programs such as stochastic in RNAstructure (t), DotKnot (d), RNALfold in ViennaRNA (l), UNAFold with ddG = 5, W = 4 (u-5-4), Fold (f), ProbKnot in RNAstructure (p), and pKiss (k) are better options. ProbKnot in RNAstructure (p) has good overall performance as measured by precision, recall, and F1. Depending on the preference of precision (F1) or recall, a combination of programs/conditions of ProbKnot with some programs with high performance in the corresponding statistic is recommended.

4.4 Discussions

RNA structures have been studied by multiple computational approaches in the last 40 years. The most widely used computational approaches are based on dynamic programming and predict both MFE and near-MFE structures for a given RNA sequence.

Other approaches use statistical sampling from the Boltzmann distribution and calculate base-pairing probabilities according to calculated free energies. In addition, novel approaches using grammar-based algorithms to get the maximum accuracy structures have emerged recently. However, structures predicted by individual programs are only partially correct.

This work focuses on finding optimal combinations of results from individual programs to improve the accuracy of structure prediction, based on the RNA topology level. As we know, RNA topology, which is the nesting, adjacency, and pseudoknotting relationships between stems, is closely related to RNA function. When comparing with the local structural features, such as number of base pairs, the global arrangement of stems is more conserved and reliable. Furthermore, pseudoknots are important structural elements in RNA functions. Some RNA structure prediction programs can predict pseudoknots, or one stem from a pseudoknot, however, there is still space for improvement, which is another challenge this work tries to tackle. For example, UNAFold predicts multiple alternative structures without pseudoknots; however, the two stems in one pseudoknot could have been predicted in two separate structures. The combined structure, which retains stem information from alternative structures, has the potential to identify the implicit pseudoknot information. Based on a global point of view, the structure combination approach has promising potential to improve structure prediction.

We have investigated the prediction performance of 24 programs/parameter combinations and correlation analysis shows that, in spite of similar methodology, different programs show considerably variability in prediction performance, and considerable differ-

ences in performance across functional families. Since the predictions made by individual programs are each only partially correct, we combine the results from multiple programs and look for combinations with improved performance. Our results have shown Pareto frontiers for each functional family, which are potential optimal solutions for precisions vs recall. For each program combination on the frontier, there is no other program or program combination that has both higher precision and higher recall. Depending on the need for higher precision or recall, program combination on the Pareto frontier represent the best possible choices for structure predictions. In the high precision region of the Pareto frontier we tend to find that the best program combinations include no more than two programs; when more than two programs are combined the level of increased noise exceeds the level of increased information. In the high recall region of the Pareto Frontier we tend to find that the best combinations include around 5 programs; this indicates that there is considerable amount of independence between the programs and that generating the most comprehensive list of true stems requires multiple approaches. Unfortunately, this high recall comes at the expense of greatly reduced precision (number of false positive predictions). F1 score, which is a balance between precision and recall, shows that combining programs that differ in algorithms yields a better overall performance (Table 4.2). For example, dynamic programming-based approaches, such as ProbKnot (RNAstructure) and grammar-based approaches, such as IPknot or CentroidFold, are complement each other. However, caution should be taken when using the F1 scores. In general, the range of precision, attends to be much larger than the range of recall. Therefore, differences in the F1 score are dominat-

ed by the differences in the precision. There is no universal best setting for precision versus recall, but the results here can be broadly used to select programs or program combinations that optimize the result required by the end-user's application, whether it is high precision, high recall, or both. This choice is simplified by the limited number of program combinations found along the Pareto frontier; other combinations need not be considered since, for any combination not on the frontier, there is always a combination on the front that has both higher precision AND higher recall.

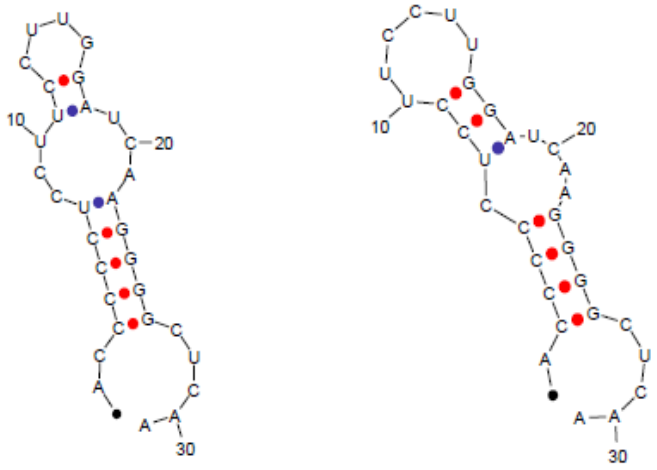


Figure 4.1 An example of alternative predicted RNA foldings sharing overlapping base-paired regions. Two alternative structures of a 31 base RNA sequence (ACCCCCUCCUUCCUUGGAUCAAGGGGCUCAA) were predicted using UNAFold. The plot shows the MFE structure (left, ΔG = -9.8 kcal/mol) and a near-MFE structure (right, ΔG = -9.5 kcal/mol). The MFE structure has two base-paired regions, 3-7:2-226, and 11-12:17-18. The near-MFE structure also has two base-paired regions, 2-5:23-26, and 7-9:16-18.

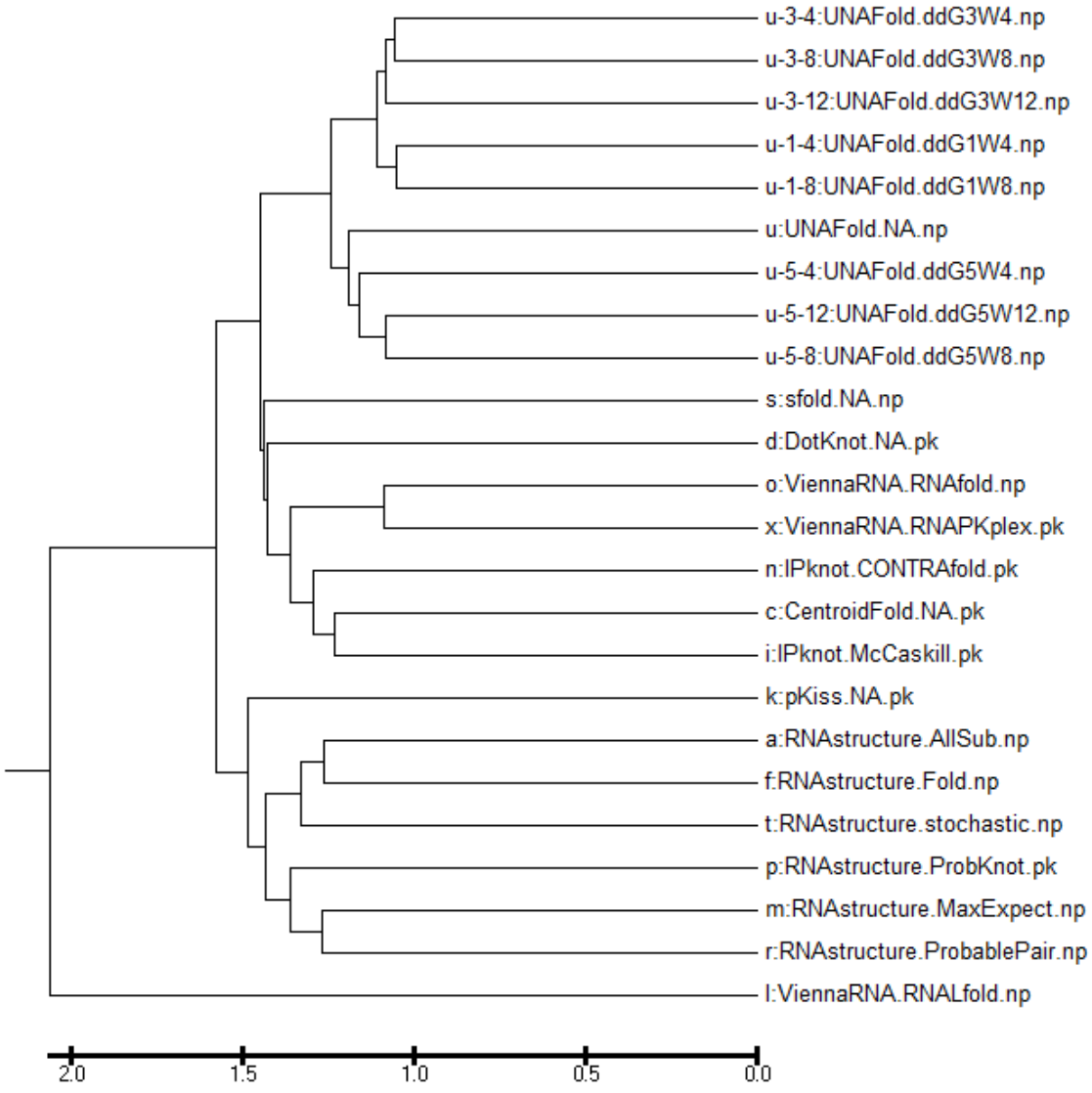


Figure 4.2 UPGMA tree of RNA structure prediction programs. The pairwise correlation between any two programs were calculated, and used to compute the distance which is the inverse of correlation. A UPGMA (Unweighted Pair Group Method with Arithmetic Mean) tree were generated from the correlation. In the UPGMA tree, the branches started from single programs; branches with least distances were joined to create a new branch, and a new distance was calculated; this process continued until all the branches were joined to a common ancestor (root of the tree).

Figure 4.3 Precision v.s. recall of all the program combinations. Each dot represents the prediction performance of one program combination for a specific curated RNA structure and is drawn with a family-specific shape. For each RNA family, the Pareto frontier is shown by shapes filled with the family-specific color and outlined in black (dots with top 10% precision and top 10% recall) or shapes filled with white and enclose outlined in the family-specific color (dots not in the top 10% precision or top 10% recall range); also, a regression line fitting all the dots of the Pareto Frontier in that family is drawn with the family-specific color. Dots not in the Pareto Frontier are drawn in grey. For the specific shape/color of each family, refer to the legend box on the top-left corner; (B) individual families, from 1 through 8, are 16S rRNA, 23S rRNA, RNase P RNA, Group II Intron, tRNA, Group I Intron, tmRNA, and 5S rRNA (the order corresponds to the order of the families plotted in (A). The non-Pareto Frontier dots within a specific family is drawn in black, with dots in other families drawn in grey.

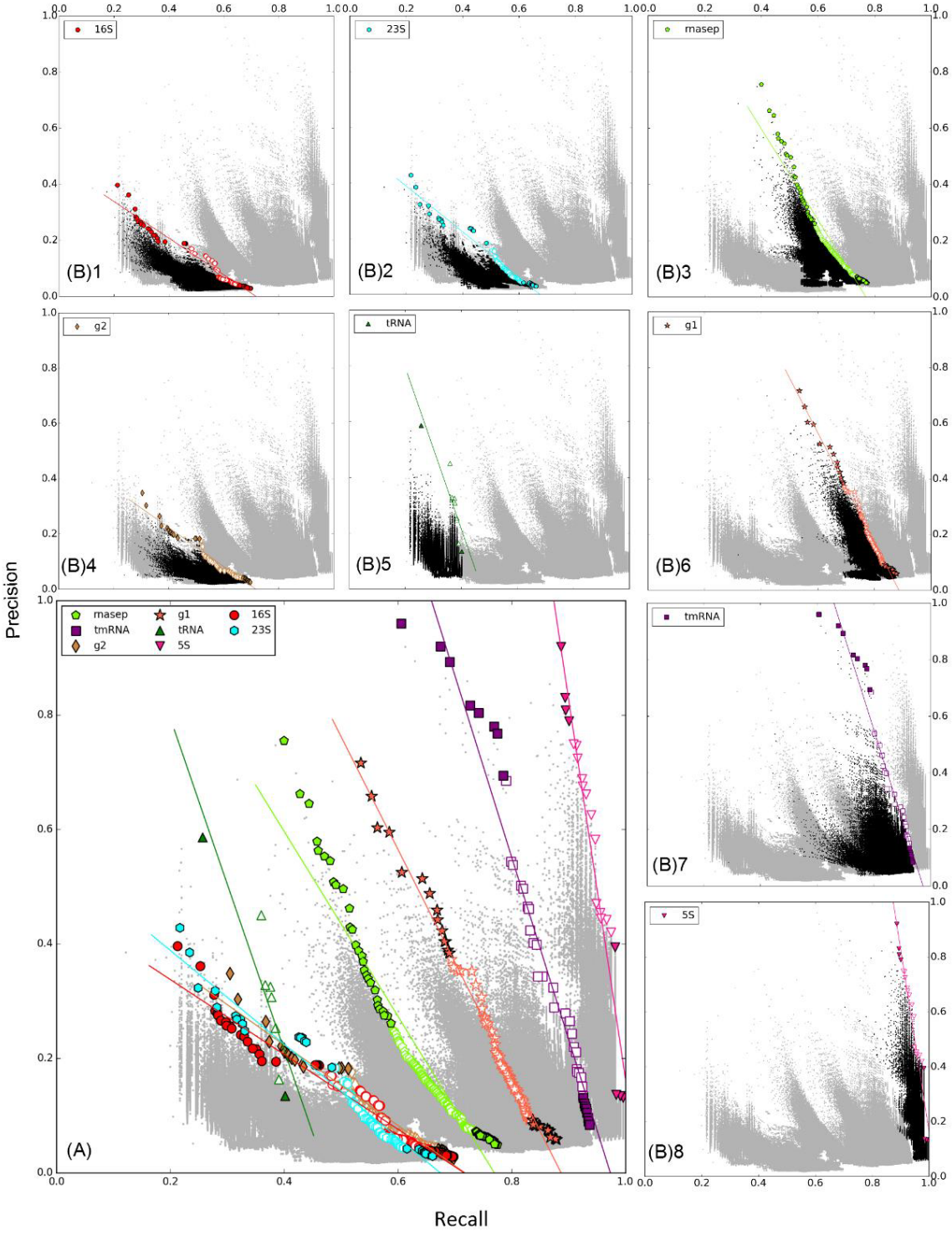

Precision
Recall
(A)
(B)1
(B)2
(B)3
(B)4
(B)5
(B)6
(B)7
(B)8
16S
23S
rnasep
g2
tRNA
g1
tmRNA
5S
rnasep
tmRNA
g2
g1
tRNA
5S
16S
23S

Figure 4.3

Table 4.1 RNA structure prediction programs tested in this study. The third column shows the abbreviation for each program used in the text and figures. For UNAFold, W = 4, 8, or 12, ΔΔG = 1, 3, or 5 kcal/mol. The abbreviation u stands for UNAFold MFE prediction, while u-1-8, stands for UNAFold suboptimal prediction at free energy range ΔΔG = 1 kcal/mol and window size W = 8.

Package	**Program**	**Code**	**Method**	**Availability of suboptimal folding(s)**	**Pseudoknot prediction**
CentroidFold	N/A	**c**	Predicts an RNA secondary structure with maximum base-pairing probability based on generalized centroid estimator that adjusts precision and recall upon user request (Hamada, et al., 2009).	No	Yes
IPknot	McCaskill	**i**	Predicts an RNA secondary structure with maximum base-pairing probability using integer programming by McCaskill model (Sato, et al., 2011).	No	Yes
	CONTRAfold	**n**	Predicts an RNA secondary structure with maximum base-pairing probability using integer programming by CONTRAfold model (Sato, et al., 2011).	No	Yes

Table 4.1 continued

RNAstructure	AllSub	**a**	Generates all possible near-minimum free energy structures (Duan, et al., 2006).	Yes	No
	Fold	**f**	Predicts the minimum free energy (MFE) suboptimal structures (Mathews, et al., 2004).	Yes	No
	MaxExpect	**m**	Generates one or several structures composed of highly probable base-pairs (Lu, et al., 2009).	No	No
	ProbKnot	**p**	Calculates the highest pairing probabilities to yield a maximum expected accuracy structure which might contains psedoknots (Bellaousov and Mathews, 2010).	No	Yes
	ProbablePair	**r**	Generates secondary structures composed of base-pairs with base-pairing probabilities that exceed a specified threshold (Mathews, 2004).	Yes	No
	stochastic	**t**	Generates a representative sample of structures by stochastic sampling from the Boltzmann ensemble (Harmanci, et al., 2009).	Yes	No

Table 4.1 continued

ViennaRNA	RNALfold	**l**	Calculates locally stable secondary structures of RNAs (Hofacker, 2003; Hofacker, 2004; Lorenz, et al., 2011).	Yes	No
	RNAfold	**o**	Computes base-pairing probabilities using partition function algorithm (55) and yields both the MFE structure and suboptimal structures (Hofacker, 2003; Hofacker, 2004; Lorenz, et al., 2011).	Yes	No
	RNAPKplex	**x**	Predicts RNA secondary structures including pseudoknots (Hofacker, 2003; Hofacker, 2004; Lorenz, et al., 2011).	No	Yes

Table 4.1 continued

UNAFold	Minimum Free Energy (MFE)	**u**	Computes minimum free energy folding and suboptimal foldings by dynamic programmingods (Markham and Zuker, 2008; Zuker, 2003; Zuker, et al., 1999).	No	No
	ddG=1, W=4	**u-1-4**		Yes	
	ddG=1, W=8	**u-1-8**			
	ddG=3, W=4	**u-3-4**			
	ddG=3, W=8	**u-3-8**			
	ddG=3, W=12	**u-3-12**			
	ddG=5, W=4	**u-5-4**			
	ddG=5, W=8	**u-5-8**			
	ddG=5, W=12	**u-5-12**			
DotKnot	N/A	**d**	Predicts RNA structures with restricted topologies H-type pseudoknots and kissing hairpins) by assembling high probability base-paired regions (Sperschneider and Datta, 2010; Sperschneider, et al., 2011).	No	Yes

Table 4.1 continued

pKiss	N/A	**k**	Predicts RNA structures with estricted topologies (H-type pseudoknots and kissing hairpins) using heuristics (Corinna Theis, 2010; Janssen and Giegerich, 2015; Reeder, et al., 2007).	Yes	Yes
sfold	N/A	**s**	Calculates the centroid structure of the Boltzmann ensemble using partition-function weighted statistical sampling (Ding, et al., 2005; Ding and Lawrence, 2003).	Yes	No

Table 4.2 F1 score of the 20 best performing programs combinations vs 8 RNA families. The second row shows the average size (number of stems) for each RNA family. For each program/condition combination, the F1 score is averaged over each of the 8 RNA families and also averaged over the 8 RNA family averages (all). Within each RNA family and also within the overall average, the F1 scores for the 327,679 program combinations are ranked from highest to lowest. For each combination, the average ranking is calculated by summing its ranking within each family and dividing by the number of families. The program combinations with the 20 highest average rankings are shown below, with the best F1 score within each family bolded.

Combination (stem_#)	16S (60.5±17.1)	23S (46.9±8.2)	5S (4.6±0.5)	g1 (13.3± 2.1)	g2 (20.0± 4.5)	Rnasep (24.3± 5.9)	tRNA (9.6± 1.0)	tmRNA (25.8± 5.9)	All (21.7±17.3)	Average Ranking
n.p	0.266±0.121	0.283±0.089	0.778±0.223	0.541±0.168	0.251±0.102	0.505±0.12	0.328±0.042	0.658±0.11	0.54±0.225	18
p	0.272±0.112	0.281±0.087	0.816±0.197	0.528±0.162	0.235±0.111	0.457±0.123	0.355±0.068	**0.774±0.076**	**0.577±0.241**	20
c.p	0.253±0.126	0.274±0.097	0.797±0.228	0.557±0.168	0.227±0.11	0.506±0.126	**0.335±0.045**	0.65±0.104	0.554±0.228	22
i.p	0.273±0.121	0.26±0.084	0.779±0.23	0.546±0.163	0.242±0.11	0.494±0.13	0.317±0.046	0.636±0.122	0.533±0.228	25
n.r	0.263±0.133	**0.289±0.115**	0.801±0.216	0.537±0.172	0.269±0.104	0.504±0.111	0.283±0.026	0.637±0.12	0.54±0.227	37
m.n	0.255±0.125	0.269±0.112	0.754±0.213	0.533±0.173	0.26±0.104	0.493±0.117	0.289±0.045	0.631±0.117	0.527±0.22	44
p.r	0.255±0.106	0.272±0.093	0.795±0.238	0.503±0.165	0.23±0.108	0.433±0.126	0.311±0.061	0.774±0.086	0.553±0.255	45
c.m	0.247±0.113	0.253±0.115	0.784±0.214	0.553±0.175	0.263±0.136	0.498±0.117	0.298±0.045	0.619±0.114	0.536±0.224	45
n.p.r	0.252±0.12	0.274±0.093	0.745±0.242	0.506±0.166	0.227±0.093	0.475±0.116	0.3±0.042	0.657±0.11	0.519±0.228	46
i	0.292±0.142	0.257±0.107	0.855±0.178	0.599±0.172	0.297±0.117	**0.517±0.133**	0.3±0.023	0.528±0.159	0.539±0.233	48
c.p.r	0.239±0.111	0.266±0.101	0.763±0.249	0.522±0.168	0.224±0.106	0.473±0.121	0.302±0.048	0.65±0.107	0.525±0.233	54
c.n.p	0.255±0.116	0.256±0.089	0.759±0.237	0.533±0.18	0.236±0.107	0.494±0.123	0.293±0.039	0.565±0.119	0.506±0.223	59
i.r	**0.275±0.126**	0.254±0.102	0.808±0.206	0.565±0.169	0.268±0.116	0.502±0.113	0.271±0.013	0.62±0.132	0.543±0.227	63
c.r	0.249±0.121	0.259±0.121	0.837±0.202	0.579±0.176	0.278±0.137	0.508±0.111	0.271±0.008	0.621±0.11	0.552±0.231	65
i.p.r	0.257±0.113	0.253±0.087	0.747±0.247	0.511±0.161	0.219±0.098	0.463±0.125	0.289±0.039	0.636±0.123	0.512±0.23	68
i.m	0.26±0.115	0.245±0.098	0.76±0.223	0.542±0.169	0.256±0.115	0.487±0.121	0.279±0.043	0.617±0.134	0.525±0.224	70
n	0.272±0.133	0.286±0.117	0.855±0.174	0.562±0.168	**0.3±0.111**	0.502±0.114	0.313±0.018	0.477±0.173	0.513±0.232	70
c	0.263±0.145	0.254±0.125	**0.893±0.157**	**0.609±0.183**	0.321±0.13	0.517±0.129	0.309±0.017	0.486±0.151	0.546±0.24	71
c.i.p	0.254±0.119	0.238±0.08	0.761±0.243	0.517±0.172	0.221±0.11	0.474±0.138	0.283±0.037	0.553±0.134	0.494±0.228	98
m.p	0.251±0.106	0.253±0.106	0.77±0.235	0.474±0.173	0.217±0.106	0.404±0.126	0.315±0.052	0.757±0.09	0.532±0.253	105

Table 4.3 Frequency of programs on the Pareto Frontier. For each statistic, the top 4 program combinations along the Pareto Frontier were identified for each family of structures, and the occurrence of individual programs summed. (Programs with only one occurrence in the high F1 or high precision lists, or with less than half of the maximum occurrence in the high recall list are not shown), represented as: program code (occurrence).

High F1	High precision	High recall
c(13)	n(11)	t(28)
i(10)	c(10)	d(20)
n(8)	i(9)	l(17)
a(6)	r(6)	u-5-4(16)
r(5)	a(5)	f(16)
k(4)	k(4)	p(16)
f(4)	p(3)	
t(4)		
p(4)		

Table 4.4 F1 scores of the 24 structure prediction programs. The F1 scores (mean±std) for RNA structures in 8 functional families plus that for all the RNA structures are shown. The best F1 score of each family is bolded. For UNAFold, fixing ΔΔG (energy threshold above MFE for suboptimal structures) and increasing W (window size, the variation between predicted structures) increases precision but reduces recall, and overall slightly increases the F1 score; in contrast, fixing W and increasing ΔΔG reduces precision and increases recall, and overall reduces the F1 score. In general, maximizing window size at a relatively low energy threshold (ddG=1, W = 8) yields the best F1 score for UNAFold.

Code	Package	Family (stem number) / Program	16S	23S	5S	g1	g2	rnasep	tRNA	tmRNA	Total/average
			60.5±17.1	46.9±8.2	4.6±0.5	13.3± 2.1	20.0± 4.5	24.3± 5.9	9.6± 1.0	25.8± 5.9	21.7±17.3
c	CentroidFold	N/A	0.263±0.145	0.254±0.125	**0.893±0.157**	**0.609±0.183**	**0.321±0.13**	**0.517±0.129**	0.309±0.017	0.486±0.151	0.546±0.24
i	IPknot	McCaskill	0.292±0.142	0.257±0.107	0.855±0.178	0.599±0.172	0.297±0.117	0.517±0.133	0.3±0.023	0.528±0.159	0.539±0.233
n		CONTRAfold	0.272±0.133	**0.286±0.117**	0.855±0.174	0.562±0.168	0.3±0.111	0.502±0.114	0.313±0.018	0.477±0.173	0.513±0.232
a	RNAstructure	AllSub	0.22±0.114	0.208±0.074	0.505±0.201	0.487±0.143	0.176±0.092	0.418±0.104	**0.384±0.112**	**0.734±0.083**	0.518±0.205
f		Fold	0.192±0.047	0.16±0.068	0.504±0.185	0.296±0.061	0.125±0.025	0.298±0.071	0.301±0.076	0.746±0.091	0.453±0.235
m		MaxExpect	0.235±0.103	0.211±0.057	0.779±0.191	0.53±0.17	0.23±0.093	0.449±0.126	0.309±0.017	0.729±0.083	0.575±0.224
p		ProbKnot	**0.272±0.112**	0.281±0.087	0.816±0.197	0.528±0.162	0.235±0.111	0.457±0.123	0.355±0.068	0.774±0.076	**0.577±0.241**
r		ProbablePair	0.231±0.091	0.185±0.074	0.767±0.18	0.518±0.144	0.237±0.128	0.426±0.111	0.32±0.013	0.736±0.081	0.576±0.22
t		stochastic	0.056±0.012	0.043±0.006	0.237±0.164	0.144±0.046	0.042±0.012	0.153±0.067	0.176±0.076	0.472±0.12	0.248±0.182
l	ViennaRNA	RNALfold	0.122±0.023	0.146±0.021	0.448±0.134	0.181±0.088	0.082±0.046	0.131±0.053	0.273±0.082	0.127±0.038	0.193±0.14
o		RNAfold	0.231±0.092	0.24±0.072	0.812±0.212	0.531±0.165	0.222±0.113	0.433±0.128	0.301±0.028	0.392±0.134	0.453±0.23
x		RNAPKplex	0.243±0.105	0.26±0.092	0.839±0.199	0.553±0.164	0.252±0.106	0.435±0.139	0.305±0.021	0.416±0.134	0.473±0.232
u	UNAFold	Minimum Free Energy (MFE)	0.124±0.049	0.108±0.022	0.54±0.184	0.334±0.124	0.117±0.066	0.329±0.076	0.286±0.033	0.21±0.064	0.296±0.17
u-1-4		ddG=1, W=4	0.202±0.089	0.206±0.053	0.596±0.209	0.472±0.148	0.165±0.067	0.412±0.109	0.293±0.063	0.298±0.103	0.37±0.186
u-1-8		ddG=1, W=8	0.204±0.091	0.206±0.057	0.628±0.217	0.475±0.143	0.177±0.073	0.426±0.099	0.293±0.033	0.307±0.111	0.382±0.191
u-3-4		ddG=3, W=4	0.187±0.078	0.165±0.048	0.477±0.18	0.394±0.157	0.144±0.064	0.343±0.086	0.289±0.092	0.224±0.078	0.305±0.16
u-3-8		ddG=3, W=8	0.196±0.082	0.178±0.047	0.556±0.186	0.429±0.142	0.161±0.067	0.389±0.088	0.293±0.033	0.266±0.092	0.346±0.173
u-3-12		ddG=3, W=12	0.206±0.086	0.188±0.045	0.625±0.195	0.456±0.143	0.17±0.072	0.409±0.103	0.305±0.02	0.29±0.094	0.372±0.187
u-5-4		ddG=5, W=4	0.096±0.037	0.092±0.037	0.401±0.181	0.26±0.101	0.1±0.057	0.261±0.071	0.252±0.075	0.149±0.053	0.221±0.139
u-5-8		ddG=5, W=8	0.124±0.046	0.122±0.037	0.548±0.182	0.367±0.124	0.124±0.063	0.347±0.072	0.286±0.033	0.227±0.071	0.304±0.172
u-5-12		ddG=5, W=12	0.15±0.057	0.148±0.038	0.625±0.195	0.427±0.138	0.142±0.08	0.394±0.09	0.305±0.02	0.27±0.085	0.349±0.191
d	DotKnot	N/A	0.251±0.085	0.236±0.065	0.646±0.233	0.559±0.151	0.195±0.09	0.42±0.132	0.292±0.017	0.4±0.166	0.433±0.209
k	pKiss	N/A	0.185±0.024	NA	0.474±0.133	0.506±0.116	0.199±0.081	0.39±0.121	0.28±0.079	0.378±0.117	0.408±0.144
s	sfold	N/A	0.214±0.119	0.205±0.089	0.77±0.212	0.543±0.172	0.198±0.091	0.469±0.142	0.296±0.023	0.354±0.133	0.451±0.234

Table 4.5 Program combinations with the top 20 precisions. The precision (mean±std) for RNA structures in 8 functional families plus that for all the RNA structures are shown. The best precision of each family is bolded.

Combination	16S	23S	5S	g1	g2	rnasep	tRNA	tmRNA	all	Average Ranking
stem_#	60.5±17.1	46.9±8.2	4.6±0.5	13.3± 2.1	20.0± 4.5	24.3± 5.9	9.6± 1.0	25.8± 5.9	21.7±17.3	NA
r	0.372±0.163	0.385±0.125	0.859±0.208	0.698±0.18	0.258±0.16	0.687±0.169	0.5±0	0.96±0.075	0.763±0.246	3
c	0.395±0.199	0.376±0.173	0.919±0.15	0.716±0.204	0.348±0.146	0.755±0.166	0.542±0.073	0.537±0.188	0.647±0.25	11
n.r	0.333±0.171	0.385±0.139	0.748±0.255	0.574±0.204	0.256±0.128	0.619±0.159	0.431±0.133	0.605±0.138	0.56±0.229	12
n	0.396±0.186	0.428±0.145	0.854±0.198	0.684±0.22	0.303±0.132	0.747±0.146	0.566±0.082	0.512±0.193	0.607±0.25	13
c.r	0.335±0.152	0.353±0.176	0.808±0.238	0.603±0.195	0.27±0.15	0.621±0.16	0.375±0.073	0.611±0.143	0.58±0.233	13
p	0.287±0.118	0.281±0.086	0.821±0.235	0.562±0.177	0.206±0.112	0.573±0.142	0.586±0.097	0.92±0.09	0.653±0.277	14
i	0.361±0.168	0.296±0.109	0.83±0.2	0.658±0.195	0.264±0.121	0.662±0.157	0.481±0.04	0.506±0.167	0.574±0.24	19
c.m	0.294±0.143	0.316±0.19	0.733±0.262	0.523±0.189	0.245±0.157	0.531±0.139	0.4±0.116	0.597±0.142	0.528±0.228	19
m.n	0.297±0.164	0.323±0.159	0.676±0.255	0.505±0.187	0.239±0.13	0.525±0.131	0.378±0.114	0.59±0.132	0.509±0.217	26
i.r	0.307±0.138	0.277±0.106	0.747±0.244	0.563±0.187	0.229±0.123	0.563±0.145	0.359±0.033	0.553±0.144	0.527±0.225	26
p.r	0.274±0.117	0.265±0.098	0.754±0.279	0.49±0.177	0.193±0.107	0.487±0.157	0.407±0.091	0.893±0.119	0.588±0.293	30
n.p	0.27±0.13	0.248±0.075	0.704±0.264	0.51±0.182	0.193±0.09	0.544±0.14	0.41±0.079	0.603±0.126	0.512±0.226	36
c.n.r	0.271±0.134	0.276±0.12	0.725±0.271	0.521±0.205	0.196±0.105	0.553±0.159	0.361±0.161	0.463±0.145	0.483±0.236	43
m	0.252±0.111	0.222±0.059	0.795±0.233	0.562±0.192	0.187±0.095	0.554±0.141	0.542±0.073	0.919±0.098	0.668±0.271	44
c.n	0.316±0.156	0.29±0.123	0.803±0.248	0.595±0.222	0.227±0.114	0.645±0.163	0.431±0.133	0.407±0.174	0.518±0.26	44
c.m.r	0.267±0.147	0.305±0.197	0.69±0.294	0.467±0.182	0.23±0.163	0.464±0.144	0.312±0.109	0.588±0.143	0.491±0.233	44
i.m	0.27±0.126	0.248±0.111	0.686±0.269	0.495±0.18	0.213±0.125	0.49±0.134	0.345±0.089	0.543±0.143	0.484±0.221	47
c.p	0.232±0.113	0.24±0.087	0.749±0.271	0.525±0.179	0.174±0.096	0.545±0.146	0.403±0.071	0.613±0.131	0.532±0.237	49
r.x	0.233±0.111	0.269±0.091	0.743±0.28	0.513±0.168	0.191±0.13	0.461±0.141	0.38±0.066	0.485±0.111	0.471±0.225	53
m.n.r	0.276±0.168	0.31±0.168	0.636±0.277	0.451±0.18	0.225±0.135	0.461±0.136	0.307±0.098	0.582±0.133	0.475±0.22	54

Table 4.6 Program combinations with the top 20 recalls. The recall (mean±std) for RNA structures in 8 functional families plus that for all the RNA structures are shown. The best recall of each family is bolded.

Combination	16S	23S	5S	g1	g2	rnasep	tRNA	tmRNA	all	Average Ranking
stem_#	60.5±17.1	46.9±8.2	4.6±0.5	13.3± 2.1	20.0± 4.5	24.3± 5.9	9.6± 1.0	25.8± 5.9	21.7±17.3	NA
a.d.f.l.m.t.u-5-4.x	0.691±0.178	0.65±0.12	0.996±0.03	0.872±0.133	0.679±0.166	0.772±0.114	0.399±0.143	0.934±0.085	0.82±0.198	3815
a.d.f.l.p.t.u-5-4.x	0.692±0.178	0.658±0.113	0.996±0.03	0.872±0.133	0.687±0.162	0.772±0.116	0.399±0.143	0.934±0.085	0.821±0.197	3951
a.d.f.k.l.t.u-5-4	0.688±0.181	0.645±0.128	0.996±0.03	0.873±0.135	0.679±0.151	0.772±0.113	0.399±0.143	0.94±0.082	**0.822±0.198**	3955
a.d.k.l.p.t.u-5-4	0.683±0.181	0.655±0.117	0.996±0.03	0.871±0.135	0.692±0.151	0.754±0.124	0.399±0.143	0.935±0.083	0.818±0.198	3975
a.d.i.k.l.t.u-5-4.x	0.685±0.176	0.653±0.12	0.996±0.03	0.87±0.129	0.686±0.15	0.75±0.116	0.399±0.143	0.939±0.079	0.818±0.197	4104
a.f.i.l.p.t.u-5-4.x	0.69±0.178	0.658±0.113	0.996±0.03	0.866±0.127	0.679±0.157	0.77±0.115	0.399±0.143	0.934±0.082	0.819±0.196	4209
a.f.k.l.p.t.u-5-4	0.686±0.182	0.652±0.116	0.996±0.03	0.867±0.133	0.681±0.148	0.768±0.112	0.399±0.143	0.936±0.083	0.819±0.197	4298
a.d.f.l.n.t.u-5-4.x	0.693±0.176	0.65±0.12	0.996±0.03	0.871±0.131	0.683±0.155	0.772±0.114	0.399±0.143	0.934±0.085	0.821±0.196	4329
a.d.k.l.s.t.u-5-4.x	0.687±0.176	0.653±0.125	0.996±0.03	0.873±0.131	0.689±0.146	0.752±0.118	0.399±0.143	0.937±0.079	0.819±0.197	4433
a.d.k.l.n.t.u-5-4.x	0.687±0.176	0.65±0.12	0.996±0.03	0.869±0.13	0.69±0.145	0.75±0.115	0.399±0.143	0.935±0.083	0.818±0.196	4483
a.d.l.n.p.t.u-5-4.x	0.687±0.175	0.658±0.113	0.996±0.03	0.867±0.131	0.691±0.158	0.756±0.125	0.399±0.143	0.928±0.087	0.817±0.197	4532
a.d.f.l.t.u-5-4.x	0.689±0.177	0.65±0.12	0.996±0.03	0.869±0.132	0.676±0.162	0.771±0.115	0.399±0.143	0.934±0.085	0.819±0.197	4546
a.d.l.p.s.t.u-5-4	0.682±0.177	0.66±0.118	0.996±0.03	0.867±0.137	0.686±0.166	0.755±0.125	0.399±0.143	0.929±0.084	0.816±0.198	4657
a.d.f.l.s.t.u-5-4.x	0.693±0.176	0.653±0.125	0.996±0.03	0.871±0.134	0.682±0.156	0.772±0.114	0.399±0.143	0.935±0.081	0.821±0.197	4680
a.d.l.n.s.t.u-5-4.x	0.688±0.174	0.653±0.125	0.996±0.03	0.868±0.131	0.687±0.152	0.754±0.119	0.399±0.143	0.929±0.084	0.816±0.196	4715
a.c.d.k.l.t.u-5-4	0.68±0.179	0.647±0.124	0.996±0.03	0.87±0.133	0.683±0.151	0.749±0.114	0.399±0.143	0.935±0.083	0.816±0.198	4732
a.d.l.m.n.t.u-5-4	0.684±0.176	0.647±0.124	0.996±0.03	0.867±0.135	0.682±0.161	0.753±0.12	0.399±0.143	0.928±0.087	0.815±0.198	4856
a.d.l.n.p.t.u-5-4	0.685±0.177	0.657±0.113	0.996±0.03	0.866±0.135	0.69±0.157	0.756±0.125	0.399±0.143	0.928±0.087	0.816±0.197	4858
a.c.d.i.l.t.u-5-4.x	0.682±0.175	0.653±0.12	0.996±0.03	0.866±0.128	0.679±0.159	0.75±0.117	0.399±0.143	0.933±0.085	0.815±0.197	4909
a.i.k.l.p.t.u-5-4	0.681±0.179	0.657±0.113	0.996±0.03	0.863±0.131	0.685±0.147	0.749±0.124	0.399±0.143	0.937±0.079	0.816±0.198	4968

Table 4.7 The precision, recall, and F1 score of the points in the Pareto Frontier in 16S rRNAs.

16S: 117 program combinations				
program combina-	precision	recall	F1	top precision/recall
n	0.396	0.213	0.277	p
i	0.361	0.253	0.298	p
i.k	0.311	0.277	0.293	p
a.k.n	0.283	0.279	0.281	p
i.n	0.278	0.283	0.28	p
a.i	0.275	0.284	0.279	p
i.k.r	0.266	0.289	0.277	p
k.p	0.263	0.298	0.279	p
k.n.p	0.257	0.299	0.276	p
i.p	0.256	0.305	0.278	p
i.k.n	0.252	0.308	0.277	p
i.k.p	0.241	0.324	0.276	
f.n	0.238	0.33	0.277	
a.f.n	0.229	0.336	0.272	
f.k.n	0.217	0.345	0.266	
f.i	0.216	0.351	0.267	
a.f.i	0.207	0.357	0.262	
a.f.i.r	0.195	0.361	0.253	
f.p.r	0.194	0.386	0.258	
n.t	0.188	0.455	0.266	
n.r.t	0.188	0.457	0.266	
a.n.r.t	0.188	0.458	0.267	
a.m.n.t	0.188	0.461	0.267	
a.m.n.r.t	0.187	0.462	0.266	
f.n.t	0.17	0.48	0.251	
a.f.m.n.t	0.169	0.483	0.25	
a.f.m.n.r.t	0.169	0.484	0.251	
c.f.m.r.t	0.154	0.485	0.234	

Table 4.7 continued

a.c.f.m.t	0.154	0.486	0.234	
f.k.m.n.t	0.151	0.494	0.231	
a.f.k.m.n.r.t	0.151	0.495	0.231	
p.r.t	0.144	0.533	0.227	
a.m.p.t	0.143	0.534	0.226	
a.m.p.r.t	0.142	0.535	0.224	
k.p.r.t	0.134	0.543	0.215	
a.k.m.p.t	0.133	0.544	0.214	
a.k.m.p.r.t	0.133	0.545	0.214	
f.m.p.t	0.127	0.557	0.207	
a.f.m.p.r.t	0.126	0.558	0.206	
f.k.p.t	0.118	0.565	0.195	
f.k.m.p.t	0.117	0.567	0.194	
f.k.m.p.r.t	0.117	0.568	0.194	
a.f.k.m.p.r.t	0.117	0.569	0.194	
d.f.p.t	0.097	0.57	0.166	
a.d.f.m.p.t	0.096	0.572	0.164	
d.f.k.m.p.t	0.091	0.578	0.157	
a.d.f.k.p.r.t	0.091	0.579	0.157	
f.n.p.t.u-3-12	0.069	0.58	0.123	
a.f.m.n.p.t.u-3-12	0.068	0.581	0.122	
f.k.n.p.t.u-3-12	0.067	0.584	0.12	
f.k.n.p.r.t.u-3-12	0.066	0.585	0.119	
d.f.m.n.t.u-3-12	0.065	0.586	0.117	
a.d.f.m.n.t.u-3-12	0.065	0.587	0.117	
f.m.n.t.u-3-4	0.064	0.588	0.115	
f.k.n.p.r.t.u-3-8	0.063	0.589	0.114	
d.f.k.m.n.r.t.u-3-12	0.063	0.591	0.114	
a.d.f.k.m.n.t.u-3-12	0.063	0.592	0.114	
d.f.k.n.t.u-3-8	0.061	0.594	0.111	
a.f.n.p.t.u-3-4	0.06	0.596	0.109	
a.f.n.p.r.t.u-3-4	0.06	0.597	0.109	

Table 4.7 continued

n.r.t.u	0.059	0.6	0.107	
m.n.t.u	0.059	0.603	0.107	
a.m.n.t.u	0.059	0.604	0.107	
a.k.m.n.t.u	0.058	0.607	0.106	
k.n.p.t.u	0.057	0.609	0.104	
a.m.n.p.r.t.u	0.057	0.61	0.104	
k.n.p.r.t.u	0.057	0.611	0.104	
a.k.n.p.r.t.u	0.057	0.612	0.104	
f.k.n.r.t.u	0.056	0.616	0.103	
f.k.m.n.t.u	0.056	0.617	0.103	
a.f.m.n.p.t.u	0.055	0.621	0.101	
f.k.n.r.s.t.u	0.054	0.624	0.099	
c.f.k.n.s.t.u	0.053	0.625	0.098	
f.k.n.p.s.t.u	0.052	0.626	0.096	
a.d.f.k.m.n.t.u-5-12	0.05	0.627	0.093	
d.f.k.n.p.t.u-5-12	0.048	0.629	0.089	
d.f.k.n.p.r.t.u-5-12	0.048	0.63	0.089	
d.f.k.m.n.r.t.u	0.047	0.631	0.087	
d.f.k.m.n.t.u	0.047	0.632	0.087	
d.f.k.n.r.s.t.u-5-12	0.047	0.633	0.088	
d.f.n.r.s.t.u	0.046	0.635	0.086	
d.f.k.n.s.t.u	0.045	0.637	0.084	
d.f.m.n.r.t.u-5-8	0.044	0.639	0.082	
f.l.m.n.t.u-5-12	0.044	0.64	0.082	
f.k.l.n.r.t.u-5-12	0.043	0.642	0.081	
d.k.l.m.n.t.u-5-12	0.043	0.645	0.081	
d.f.k.l.t.u-5-12	0.042	0.647	0.079	
a.d.f.k.l.r.t.u-5-12	0.042	0.648	0.079	
d.f.l.m.n.t.u-5-12	0.042	0.65	0.079	
d.f.k.l.n.t.u-5-12	0.041	0.652	0.077	
a.d.f.k.l.n.t.u-5-12	0.041	0.653	0.077	
d.f.k.l.m.n.t.u-5-12	0.041	0.654	0.077	

Table 4.7 continued

d.f.k.l.n.p.t.u-5-12	0.04	0.655	0.075	
d.f.k.l.n.s.t.u-5-12	0.039	0.657	0.074	
d.f.l.m.n.t.u	0.038	0.658	0.072	
d.f.k.l.n.r.t.u	0.038	0.659	0.072	
d.k.l.m.n.r.t.u-5-8	0.038	0.66	0.072	
d.f.l.m.n.t.u-5-8	0.037	0.666	0.07	
d.f.k.l.n.t.u-5-8	0.037	0.667	0.07	
f.t	0.037	0.67	0.07	
a.f.t	0.036	0.672	0.068	
a.f.r.t	0.036	0.675	0.068	
f.m.r.t	0.036	0.677	0.068	
a.f.m.r.t	0.035	0.678	0.067	
a.d.f.k.n.s.t.u-5-4	0.032	0.679	0.061	
d.f.k.n.s.t.u-5-4	0.032	0.68	0.061	
d.f.k.n.o.p.t.u-5-4	0.032	0.681	0.061	r
d.f.k.n.o.s.t.u-5-4	0.031	0.682	0.059	r
d.f.l.t.u-5-4	0.03	0.685	0.057	r
f.k.l.m.n.r.t.u-5-4	0.03	0.686	0.057	r
a.f.l.m.n.t.u-5-4	0.03	0.687	0.057	r
f.k.l.m.n.t.u-5-4	0.03	0.689	0.057	r
d.f.l.m.n.t.u-5-4.x	0.029	0.694	0.056	r
d.f.k.l.n.p.t.u-5-4	0.029	0.695	0.056	r
d.f.l.n.p.t.u-5-4.x	0.028	0.696	0.054	r
d.f.k.l.n.p.t.u-5-4.x	0.028	0.697	0.054	r
d.f.k.l.n.s.t.u-5-4.x	0.028	0.698	0.054	r

Table 4.8 The precision, recall, and F1 score of the points in the Pareto Frontier in 23S rRNAs.

23S: 83 program combinations				
program combina-	precision	recall	F1	top precision/recall
k.n	0.428	0.217	0.288	p
n.r	0.385	0.234	0.291	p
m.n	0.323	0.249	0.281	p
a.k.n	0.318	0.279	0.297	p
a.k.m.n	0.289	0.282	0.285	p
c.f	0.274	0.315	0.293	p
c.f.m	0.268	0.317	0.29	p
f.n	0.268	0.324	0.293	p
a.f.k.n	0.26	0.327	0.29	
k.n.p	0.248	0.331	0.284	
c.k.t	0.237	0.426	0.305	
c.f.t	0.236	0.429	0.304	
a.c.f.k.t	0.236	0.432	0.305	
a.n.t	0.23	0.436	0.301	
a.f.n.r.t	0.228	0.439	0.3	
k.m.p.r.t	0.184	0.484	0.267	
a.p.r.t	0.184	0.487	0.267	
a.f.k.m.p.r.t	0.183	0.49	0.266	
c.k.p.r.t	0.159	0.505	0.242	
a.c.k.p.t	0.159	0.508	0.242	
a.c.f.m.p.t	0.157	0.511	0.24	
a.f.k.n.p.r.t	0.155	0.512	0.238	
a.i.p.r.t	0.136	0.515	0.215	
a.f.i.m.p.t	0.136	0.518	0.215	
a.d.k.p.t	0.123	0.52	0.199	
a.d.f.m.p.t	0.122	0.523	0.198	
a.f.i.k.n.p.r.t	0.119	0.524	0.194	
d.n.p.r.t	0.112	0.53	0.185	
a.d.m.n.p.r.t	0.112	0.533	0.185	

Table 4.8 continued

a.d.f.k.n.p.r.t	0.111	0.536	0.184	
a.d.f.i.k.p.t	0.101	0.539	0.17	
c.m.p.r.t.u-1-4	0.095	0.542	0.162	
a.c.k.p.t.u-1-4	0.094	0.545	0.16	
a.c.f.k.m.p.t.u-1-4	0.094	0.547	0.16	
a.f.n.p.t.u-1-4	0.093	0.548	0.159	
a.i.k.p.t.u-1-4	0.09	0.551	0.155	
a.f.i.k.p.r.t.u-1-4	0.089	0.553	0.153	
a.d.f.i.k.m.t.u-1-4	0.087	0.554	0.15	
a.f.i.m.n.p.t.u-1-4	0.084	0.556	0.146	
a.c.d.f.k.m.p.t.u-1-4	0.082	0.557	0.143	
a.d.i.k.m.p.t.u-1-4	0.08	0.561	0.14	
a.d.f.i.k.p.r.t.u-1-4	0.079	0.564	0.139	
a.d.f.i.k.p.s.t.u-1-4	0.071	0.567	0.126	
a.d.f.i.k.p.t.u-1-4.x	0.07	0.569	0.125	
a.d.f.i.k.o.p.t.u-1-4	0.066	0.572	0.118	
a.c.p.r.t.u-5-12	0.065	0.573	0.117	
a.c.f.k.p.t.u-5-12	0.064	0.575	0.115	
m.t.u.x	0.061	0.58	0.11	
a.l.n.p.r.t.u-1-4	0.061	0.582	0.11	
a.k.t.u.x	0.061	0.583	0.11	
a.f.k.m.r.t.u.x	0.06	0.586	0.109	
a.f.o.t.u	0.059	0.588	0.107	
a.k.m.p.t.u.x	0.059	0.589	0.107	
a.f.k.m.p.r.t.u.x	0.058	0.591	0.106	
a.f.i.m.p.t.u.x	0.057	0.594	0.104	
a.f.p.r.s.t.u.x	0.056	0.595	0.102	
a.f.i.k.o.p.t.u	0.056	0.596	0.102	
a.d.f.i.p.t.u.x	0.055	0.597	0.101	
a.d.i.k.l.p.t.u-1-4.x	0.052	0.598	0.096	
a.f.i.k.p.t.u-5-8	0.05	0.6	0.092	
a.d.i.k.l.o.p.t.u-1-4	0.05	0.601	0.092	

Table 4.8 continued

a.d.i.k.l.p.s.t.u-1-4.x	0.048	0.602	0.089	
a.c.l.n.p.t.u-5-12	0.046	0.604	0.085	
a.k.l.p.s.t.u-5-12	0.045	0.605	0.084	
a.c.d.l.p.t.u-5-12	0.045	0.606	0.084	
a.d.i.l.p.t.u-5-12	0.044	0.609	0.082	
a.d.k.l.p.s.t.u-5-12	0.043	0.61	0.08	
a.l.m.p.t.u-5-8	0.042	0.616	0.079	
a.k.m.p.t.u-5-4	0.041	0.637	0.077	
a.f.k.p.t.u-5-4	0.041	0.64	0.077	
a.c.f.p.t.u-5-4	0.04	0.642	0.075	
a.f.m.p.t.u-5-4.x	0.038	0.643	0.072	
a.f.i.p.t.u-5-4	0.038	0.645	0.072	
a.f.i.k.m.p.t.u-5-4.x	0.036	0.646	0.068	
a.f.i.k.p.s.t.u-5-4	0.035	0.648	0.066	
a.f.i.p.s.t.u-5-4.x	0.033	0.649	0.063	p
a.k.l.m.p.r.t.u-5-4	0.033	0.652	0.063	p
a.c.k.l.p.r.t.u-5-4	0.032	0.654	0.061	p
a.k.l.p.t.u-5-4.x	0.032	0.655	0.061	p
a.i.l.m.p.t.u-5-4	0.032	0.657	0.061	p
a.i.k.l.p.t.u-5-4.x	0.031	0.658	0.059	p
a.d.l.p.s.t.u-5-4	0.03	0.66	0.057	p
i.l.o.p.s.t.u-5-4	0.029	0.661	0.056	p

Table 4.9 The precision, recall, and F1 score of the points in the Pareto Frontier in 5S rRNAs.

5S: 21 program combinations				
program combina-	precision	recall	F1	top precision/recall
c	0.919	0.887	0.903	p
i	0.83	0.894	0.861	p
c.r	0.808	0.895	0.849	
c.i	0.789	0.901	0.841	
i.n	0.749	0.909	0.821	
p.x	0.746	0.914	0.821	
c.i.n	0.724	0.916	0.809	
i.n.r	0.688	0.924	0.789	
i.r.x	0.675	0.925	0.78	
i.n.x	0.661	0.931	0.773	
i.n.r.x	0.625	0.94	0.751	
i.n.p.r.x	0.582	0.947	0.721	
i.u-3-8	0.47	0.949	0.629	
c.i.u-3-8	0.454	0.956	0.616	
i.n.u-3-8	0.444	0.957	0.607	
i.u-3-8.x	0.442	0.967	0.607	
i.n.u-3-8.x	0.42	0.975	0.587	
i.n.p.u-3-8.x	0.394	0.982	0.562	
c.t.u-3-8	0.136	0.984	0.239	
p.t.u-3-8	0.134	0.991	0.236	r
c.p.t.u-3-8.x	0.131	0.996	0.232	r

Table 4.10 The precision, recall, and F1 score of the points in the Pareto Frontier in Group I Introns.

g1: 78 program combinations				
program combina-	precision	recall	F1	top precision/recall
c	0.716	0.535	0.612	p
i	0.658	0.554	0.602	p
c.r	0.603	0.564	0.583	p
c.n	0.595	0.585	0.59	p
c.p	0.525	0.607	0.563	p
c.d	0.514	0.643	0.571	p
d.i	0.488	0.656	0.56	p
c.d.n	0.459	0.669	0.544	
c.d.i	0.442	0.67	0.533	
c.d.p	0.423	0.678	0.521	
d.i.m	0.404	0.682	0.507	
d.i.p	0.404	0.684	0.508	
d.i.s	0.389	0.688	0.497	
c.d.n.x	0.384	0.69	0.493	
c.d.n.p	0.384	0.691	0.494	
c.d.i.p	0.373	0.694	0.485	
a.d	0.359	0.698	0.474	
k.p	0.354	0.704	0.471	
k.s	0.353	0.71	0.472	
d.k	0.352	0.73	0.475	
d.k.r	0.328	0.737	0.454	
d.k.p	0.307	0.747	0.435	
d.k.p.r	0.287	0.748	0.415	
d.k.p.s	0.272	0.752	0.4	
c.d.k.n.s	0.27	0.753	0.397	
d.i.k.n.s	0.262	0.754	0.389	
a.d.k	0.26	0.758	0.387	
a.d.k.r	0.247	0.765	0.373	
a.d.k.p	0.235	0.77	0.36	

Table 4.10 continued

a.d.k.p.r	0.224	0.771	0.347	
c.d.k.n.s.u-3-12	0.21	0.772	0.33	
d.i.k.n.s.u-3-12	0.206	0.775	0.325	
c.d.n.p.u-5-8	0.202	0.779	0.321	
c.d.i.n.p.u-5-8	0.19	0.78	0.306	
c.d.k.u-5-8	0.188	0.783	0.303	
c.d.n.p.s.u-5-8	0.183	0.785	0.297	
a.c.d.n.u-5-8	0.182	0.788	0.296	
c.d.k.n.u-5-8	0.18	0.79	0.293	
c.d.k.p.u-5-8	0.175	0.792	0.287	
a.c.d.n.p.u-5-8	0.17	0.793	0.28	
c.d.k.n.p.u-5-8	0.168	0.796	0.277	
c.d.k.n.s.u-5-8	0.166	0.797	0.275	
c.d.k.p.s.u-5-8	0.161	0.798	0.268	
c.d.k.p.r.s.u-5-8	0.156	0.799	0.261	
c.d.k.n.p.s.u-5-8	0.155	0.802	0.26	
d.f.k.r	0.149	0.804	0.251	
a.c.d.k.n.p.u-5-8	0.146	0.805	0.247	
d.f.k.p	0.145	0.809	0.246	
d.f.k.p.r	0.141	0.811	0.24	
d.f.k.m.p.r	0.133	0.812	0.229	
a.d.f.k.r	0.131	0.816	0.226	
a.d.f.k.p	0.127	0.819	0.22	
a.d.f.k.p.r	0.124	0.82	0.215	
c.d.k.n.p.s.u-5-4	0.118	0.822	0.206	
c.d.k.n.p.s.u-5-4.x	0.112	0.823	0.197	
a.c.d.k.p.s.u-5-4	0.109	0.824	0.193	
a.c.d.k.p.s.u-5-4.x	0.104	0.825	0.185	
c.d.l.p.s.u-5-4	0.1	0.826	0.178	
d.i.l.p.s.u-5-4	0.099	0.827	0.177	
c.d.l.n.p.s.u-5-4	0.098	0.83	0.175	
c.d.i.l.p.s.u-5-4.x	0.094	0.831	0.169	

Table 4.10 continued

c.d.k.l.n.p.u-5-4	0.093	0.832	0.167	
c.d.k.t	0.093	0.833	0.167	
c.d.k.m.t	0.092	0.837	0.166	
c.d.k.m.n.t	0.084	0.838	0.153	
k.p.s.t	0.083	0.841	0.151	
k.m.s.t	0.083	0.842	0.151	
c.d.k.s.t	0.083	0.843	0.151	
d.k.s.t	0.083	0.848	0.151	
d.k.p.s.t	0.081	0.85	0.148	
d.k.t	0.076	0.863	0.14	
d.k.r.t	0.074	0.864	0.136	r
d.k.p.t	0.073	0.866	0.135	r
d.k.m.t	0.073	0.867	0.135	r
d.f.m.t	0.062	0.868	0.116	r
d.f.k.t	0.06	0.876	0.112	r
d.f.k.r.t	0.059	0.877	0.111	r
d.f.k.p.r.t	0.058	0.879	0.109	r

Table 4.11 The precision, recall, and F1 score of the points in the Pareto Frontier in Group II Introns.

g2: 86 program combinations				
program combina-	precision	recall	F1	top precision/recall
c	0.348	0.305	0.325	p
n	0.303	0.319	0.311	p
i	0.264	0.368	0.307	p
i.r	0.229	0.374	0.284	p
c.f	0.221	0.395	0.283	p
a.c.f	0.216	0.4	0.281	p
c.f.r	0.211	0.401	0.277	p
a.c.f.r	0.207	0.406	0.274	p
c.f.m	0.205	0.408	0.273	
a.c.f.m	0.201	0.413	0.27	
c.f.m.r	0.2	0.414	0.27	
a.c.f.m.r	0.197	0.419	0.268	
f.i	0.189	0.432	0.263	
a.f.i	0.185	0.434	0.259	
c.t	0.183	0.497	0.268	
a.c.t	0.183	0.501	0.268	
a.c.f.t	0.182	0.513	0.269	
c.f.m.t	0.179	0.514	0.266	
a.c.f.m.t	0.178	0.517	0.265	
a.c.f.m.r.t	0.176	0.519	0.263	
f.i.m.t	0.155	0.521	0.239	
a.f.i.m.t	0.155	0.523	0.239	
a.f.i.m.r.t	0.154	0.525	0.238	
a.c.f.k.m.t	0.127	0.526	0.205	
a.c.f.k.m.r.t	0.125	0.528	0.202	
c.f.t.x	0.118	0.532	0.193	
a.c.f.t.x	0.117	0.535	0.192	
c.f.m.t.x	0.114	0.541	0.188	
c.f.m.r.t.x	0.113	0.543	0.187	

Table 4.11 continued

a.c.f.m.r.t.x	0.112	0.545	0.186	
a.f.i.t.x	0.104	0.547	0.175	
f.i.m.t.x	0.102	0.553	0.172	
a.f.i.m.t.x	0.102	0.556	0.172	
a.f.i.m.r.t.x	0.101	0.558	0.171	
k.n.p.s.u	0.096	0.56	0.164	
i.k.p.u	0.094	0.568	0.161	
i.k.p.s.u	0.09	0.572	0.156	
a.i.k.p.u	0.085	0.573	0.148	
a.i.k.p.s.u	0.081	0.577	0.142	
i.k.n.p.u	0.075	0.582	0.133	
f.k.n.p.u	0.072	0.585	0.128	
i.k.n.p.s.u	0.071	0.586	0.127	
a.f.k.n.p.u	0.071	0.587	0.127	
a.i.k.n.p.u	0.07	0.588	0.125	
f.k.n.p.s.u	0.068	0.589	0.122	
a.n.t.u	0.067	0.601	0.121	
f.n.t.u	0.066	0.603	0.119	
a.f.n.t.u	0.066	0.606	0.119	
n.p.t.u	0.065	0.618	0.118	
a.f.n.p.t.u	0.064	0.623	0.116	
a.c.f.n.p.t.u	0.062	0.624	0.113	
a.k.n.p.t.u	0.061	0.625	0.111	
a.f.k.n.p.t.u	0.06	0.63	0.11	
c.f.k.n.p.t.u	0.059	0.631	0.108	
f.k.n.p.s.t.u	0.057	0.634	0.105	
n.p.t.u-5-8	0.051	0.635	0.094	
a.f.n.p.t.u-5-8	0.05	0.64	0.093	
a.f.n.p.r.t.u-5-8	0.049	0.642	0.091	
c.f.n.p.r.t.u-5-8	0.047	0.643	0.088	
k.n.p.r.t.u-5-8	0.046	0.644	0.086	
f.k.n.p.t.u-5-8	0.046	0.647	0.086	

Table 4.11 continued

f.k.n.p.r.t.u-5-8	0.045	0.649	0.084	
a.p.t.u-5-4	0.044	0.651	0.082	
f.p.t.u-5-4	0.044	0.656	0.082	
c.f.p.t.u-5-4	0.043	0.66	0.081	
k.p.t.u-5-4	0.042	0.661	0.079	
f.k.p.t.u-5-4	0.042	0.666	0.079	
c.f.k.p.t.u-5-4	0.041	0.667	0.077	
f.n.p.t.u-5-4	0.041	0.67	0.077	
c.f.n.p.t.u-5-4	0.04	0.671	0.075	
f.n.p.r.t.u-5-4	0.04	0.672	0.076	
f.k.n.p.t.u-5-4	0.039	0.677	0.074	
c.f.k.n.p.t.u-5-4	0.038	0.678	0.072	
f.k.n.p.r.t.u-5-4	0.038	0.679	0.072	
f.k.n.p.s.t.u-5-4	0.037	0.681	0.07	
f.k.n.p.r.t.u-5-4.x	0.035	0.684	0.067	
f.k.n.p.s.t.u-5-4.x	0.033	0.686	0.063	
d.f.k.n.p.t.u-5-4.x	0.031	0.687	0.059	
k.l.n.p.t.u-5-4	0.028	0.69	0.054	r
k.l.n.p.r.t.u-5-4	0.028	0.691	0.054	r
d.k.l.p.t.u-5-4	0.027	0.692	0.052	r
k.l.n.p.t.u-5-4.x	0.027	0.693	0.052	r
d.k.l.p.r.t.u-5-4	0.027	0.694	0.052	r
d.k.l.p.s.t.u-5-4	0.026	0.696	0.05	r
k.l.n.p.s.t.u-5-4.x	0.026	0.697	0.05	r
d.k.l.n.p.t.u-5-4.x	0.025	0.698	0.048	r

Table 4.12 The precision, recall, and F1 score of the points in the Pareto Frontier in RNase P RNAs.

Rnasep: 145 program combinations				
program combina-	precision	recall	F1	top precision/recall
c	0.755	0.4	0.523	p
i	0.662	0.428	0.52	p
c.n	0.645	0.444	0.526	p
i.n	0.579	0.458	0.511	p
i.r	0.563	0.46	0.506	p
c.n.r	0.553	0.471	0.509	p
c.p	0.545	0.481	0.511	p
i.n.r	0.508	0.487	0.497	p
i.p	0.503	0.491	0.497	p
c.n.p	0.496	0.504	0.5	p
i.n.p	0.462	0.515	0.487	p
n.p.s	0.429	0.516	0.468	p
c.i.n.p	0.425	0.52	0.468	p
c.n.p.s	0.398	0.526	0.453	p
a.c.n	0.388	0.531	0.448	
i.n.p.s	0.379	0.537	0.444	
a.i.n	0.369	0.539	0.438	
c.i.n.p.s	0.354	0.541	0.428	
a.c.i.n	0.345	0.545	0.423	
a.i.n.r	0.339	0.548	0.419	
a.c.n.p	0.332	0.552	0.415	
a.i.n.p	0.319	0.56	0.406	
a.c.i.n.p	0.302	0.564	0.393	
i.m.n.u-1-4	0.292	0.565	0.385	
a.i.u-1-8	0.286	0.568	0.38	
a.c.n.u-1-8	0.282	0.574	0.378	
i.n.p.s.u-1-8	0.275	0.576	0.372	
a.i.n.u-1-8	0.274	0.582	0.373	
a.i.n.u-1-4	0.261	0.588	0.362	

Table 4.12 continued

a.c.i.n.u-1-4	0.249	0.592	0.351	
a.i.n.r.u-1-4	0.247	0.593	0.349	
a.i.n.p.u-1-8	0.246	0.594	0.348	
a.i.n.u-5-12	0.242	0.597	0.344	
a.i.n.p.u-1-4	0.236	0.599	0.339	
a.c.i.n.u-5-12	0.231	0.601	0.334	
a.i.n.r.u-5-12	0.229	0.602	0.332	
a.c.i.n.p.u-1-4	0.226	0.603	0.329	
a.i.n.p.u-5-12	0.22	0.608	0.323	
a.c.i.n.p.u-5-12	0.212	0.612	0.315	
c.f	0.206	0.613	0.308	
a.c.i.n.p.r.u-5-12	0.202	0.616	0.304	
c.f.n	0.2	0.62	0.302	
f.i.n	0.196	0.624	0.298	
c.f.n.r	0.192	0.627	0.294	
c.f.i.n	0.19	0.628	0.292	
c.f.i.r	0.188	0.63	0.29	
f.i.n.r	0.188	0.631	0.29	
c.f.n.p	0.186	0.632	0.287	
c.f.i.n.r	0.183	0.634	0.284	
c.f.i.m	0.182	0.636	0.283	
f.i.m.n	0.182	0.637	0.283	
d.f.i.n	0.178	0.638	0.278	
c.f.i.m.n	0.177	0.64	0.277	
c.d.f.i.n	0.173	0.641	0.272	
c.f.i.n.p.r	0.171	0.642	0.27	
d.f.i.n.r	0.171	0.643	0.27	
c.f.i.m.n.r	0.17	0.644	0.269	
c.f.n.u-1-4	0.167	0.645	0.265	
c.d.f.i.n.r	0.167	0.646	0.265	
d.f.i.m.n	0.166	0.647	0.264	
c.d.f.m.n.r	0.163	0.648	0.26	

Table 4.12 continued

c.d.f.i.n.s	0.162	0.649	0.259	
c.d.f.i.m.n	0.162	0.651	0.259	
f.i.n.r.u-1-4	0.159	0.653	0.256	
c.f.n.p.u-1-4	0.157	0.654	0.253	
c.f.m.n.u-1-4	0.157	0.655	0.253	
c.f.i.n.r.u-1-4	0.155	0.656	0.251	
c.d.f.n.u-1-4	0.154	0.657	0.25	
c.f.m.n.r.u-1-4	0.152	0.658	0.247	
c.f.i.m.n.u-1-4	0.15	0.66	0.244	
c.d.f.n.r.u-1-4	0.15	0.662	0.245	
c.d.f.i.r.u-1-4	0.147	0.663	0.241	
c.f.i.m.n.r.u-1-4	0.146	0.664	0.239	
c.d.f.i.n.r.u-1-4	0.144	0.667	0.237	
d.f.i.n.r.u-5-12	0.141	0.668	0.233	
c.d.f.i.m.n.u-1-4	0.14	0.669	0.232	
c.d.f.i.n.r.u-5-12	0.138	0.672	0.229	
c.d.f.i.m.n.u-5-12	0.135	0.674	0.225	
a.c.d.f.i.n.u-1-4	0.133	0.675	0.222	
a.d.f.i.n.u-5-12	0.131	0.676	0.219	
a.c.d.f.i.n.u-5-12	0.129	0.679	0.217	
c.d.f.i.n.r.u-3-4	0.125	0.681	0.211	
c.d.f.n.p.r.u-3-4	0.123	0.682	0.208	
c.d.f.i.r.u	0.123	0.683	0.208	
c.d.f.i.n.r.u	0.121	0.686	0.206	
c.d.f.m.n.r.u	0.119	0.687	0.203	
c.d.f.i.m.n.u	0.118	0.688	0.201	
a.c.d.f.n.r.u	0.114	0.69	0.196	
a.c.d.f.i.n.u	0.113	0.691	0.194	
a.c.d.f.i.r.u	0.112	0.692	0.193	
c.d.f.u-5-4	0.107	0.694	0.185	
c.d.f.r.u-5-4	0.105	0.699	0.183	
c.d.f.n.r.u-5-4	0.103	0.702	0.18	

Table 4.12 continued

d.f.i.n.r.u-5-4	0.102	0.703	0.178	
c.d.f.i.n.r.u-5-4	0.101	0.706	0.177	
c.d.f.n.p.r.u-5-4	0.1	0.707	0.175	
c.d.f.n.p.s.u-5-4	0.098	0.708	0.172	
a.c.d.f.n.r.u-5-4	0.096	0.711	0.169	
a.c.d.f.i.r.u-5-4	0.095	0.712	0.168	
a.c.d.f.k.r.u-5-4	0.088	0.713	0.157	
d.f.l.n.r.u-5-4	0.083	0.716	0.149	
c.d.f.l.n.r.u-5-4	0.082	0.717	0.147	
d.f.i.l.n.r.u-5-4	0.081	0.719	0.146	
d.f.i.l.m.n.u-5-4	0.08	0.72	0.144	
d.f.l.n.p.s.u-5-4	0.079	0.721	0.142	
d.f.l.m.n.r.u-5-4.x	0.078	0.722	0.141	
a.d.f.l.n.r.u-5-4	0.078	0.724	0.141	
p.r.t.u	0.077	0.726	0.139	
m.p.t.u	0.076	0.727	0.138	
d.m.t.u	0.075	0.729	0.136	
d.p.t.u	0.075	0.731	0.136	
c.d.p.t.u	0.074	0.732	0.134	
d.p.r.t.u	0.074	0.733	0.134	
d.i.m.t.u	0.074	0.734	0.134	
c.d.i.m.t.u	0.073	0.735	0.133	
d.i.m.r.t.u	0.072	0.736	0.131	
d.p.r.s.t.u	0.072	0.737	0.131	
c.d.i.p.r.t.u	0.072	0.738	0.131	
c.d.m.p.r.t.u	0.071	0.739	0.13	
d.i.m.p.r.t.u	0.071	0.74	0.13	
d.n.p.t.u-5-4	0.067	0.742	0.123	
d.m.n.r.t.u-5-4	0.066	0.743	0.121	
d.n.p.r.t.u-5-4	0.066	0.744	0.121	
d.f.i.n.r.t.u-5-12	0.065	0.745	0.12	
f.r.t.u	0.065	0.747	0.12	

Table 4.12 continued

f.n.r.t.u	0.064	0.748	0.118	
f.m.t.u	0.064	0.749	0.118	
f.i.r.t.u	0.064	0.751	0.118	
f.n.p.r.t.u	0.063	0.753	0.116	
c.d.f.r.t.u	0.063	0.754	0.116	
d.f.m.r.t.u	0.062	0.756	0.115	
c.f.i.m.r.t.u	0.062	0.757	0.115	r
d.f.i.n.r.t.u	0.062	0.758	0.115	r
d.f.i.m.r.t.u	0.061	0.76	0.113	r
f.i.m.t.u-5-4	0.058	0.761	0.108	r
f.i.n.p.r.t.u-5-4	0.057	0.764	0.106	r
c.f.i.p.r.t.u-5-4	0.057	0.765	0.106	r
d.f.i.r.t.u-5-4	0.057	0.766	0.106	r
c.d.f.i.p.t.u-5-4	0.056	0.767	0.104	r
d.f.i.m.r.t.u-5-4	0.056	0.769	0.104	r
c.f.l.p.r.t.u-5-4	0.05	0.77	0.094	r
c.d.f.i.l.t.u-5-4	0.05	0.772	0.094	r
c.d.f.l.m.t.u-5-4	0.05	0.773	0.094	r
d.f.i.l.r.t.u-5-4	0.05	0.774	0.094	r
d.f.i.l.m.t.u-5-4	0.049	0.775	0.092	r

Table 4.13 The precision, recall, and F1 score of the points in the Pareto Frontier in tmRNAs.

tmRNA: 41 program combinations				
program combina-	precision	recall	F1	top precision/recall
r	0.96	0.606	0.743	p
p	0.92	0.675	0.779	p
p.r	0.893	0.691	0.779	p
f	0.817	0.727	0.769	p
f.r	0.804	0.742	0.772	
f.p	0.78	0.769	0.774	
f.p.r	0.768	0.775	0.771	
a.f.p	0.694	0.785	0.737	
a.f.p.r	0.685	0.79	0.734	
f.n.p	0.544	0.799	0.647	
f.n.p.r	0.538	0.802	0.644	
f.i.p	0.502	0.816	0.622	
f.i.p.r	0.497	0.82	0.619	
a.f.i.p	0.465	0.828	0.596	
a.f.i.p.r	0.461	0.83	0.593	
d.f.p.r	0.424	0.832	0.562	
a.d.f.p	0.402	0.838	0.543	
a.d.f.p.r	0.399	0.844	0.542	
a.c.d.f.p	0.343	0.846	0.488	
t	0.343	0.854	0.489	
f.t	0.324	0.873	0.473	
n.t	0.289	0.874	0.434	
i.t	0.279	0.894	0.425	
f.i.t	0.265	0.902	0.41	
d.f.t	0.243	0.906	0.383	
c.d.f.t	0.22	0.909	0.354	
d.f.i.t	0.211	0.913	0.343	
i.k.t	0.181	0.914	0.302	
d.f.i.s.t	0.177	0.915	0.297	

Table 4.13 continued

f.i.k.t	0.175	0.918	0.294	
d.f.k.t	0.167	0.922	0.283	
d.f.i.k.r.t	0.151	0.924	0.26	
d.f.i.k.s.t	0.134	0.925	0.234	
d.f.k.t.u-1-8	0.131	0.926	0.23	
d.f.i.k.r.t.u-1-8	0.121	0.928	0.214	
f.i.k.r.t.u-3-8	0.116	0.929	0.206	
d.f.k.r.t.u-3-8	0.112	0.931	0.2	
d.f.i.k.p.t.u-3-8	0.105	0.932	0.189	r
d.f.i.k.s.t.u-3-8	0.096	0.934	0.174	r
d.f.k.t.u	0.09	0.935	0.164	r
d.f.k.s.t.u	0.084	0.937	0.154	r

Table 4.14 The precision, recall, and F1 score of the points in the Pareto Frontier in tRNAs.

tRNA: 8 program combinations				
program combina-	precision	recall	F1	top precision/recall
p	0.586	0.257	0.357	p
a	0.45	0.36	0.4	
a.c	0.328	0.367	0.346	
a.n	0.325	0.376	0.349	
a.u-1-4	0.307	0.378	0.339	
a.c.u-1-4	0.253	0.385	0.305	
a.l.n.u-3-4	0.163	0.391	0.23	
a.t	0.134	0.402	0.201	r

CHAPTER 5. COMPUTATIONAL DESIGN OF DECOY RNA STRUCTURES USING A GRAPHICAL APPROACH

5.1 Introduction

Decoy structures, as their name indicates, are non-biological structures that have similar topological characteristics to natural molecules, but have no real biological functions. Compared to random structures which have little similarity and are easy to be discriminated from biological structures, decoy structures make better experimental controls because they share similarity with biological structures. Real biological structures often conserve some patterns that are not found in random structures. In the studies of protein structures, decoy protein structures have been used for evaluation of energy functions in protein structure prediction since the 1990s (160-165). In exploration of the RNA world, the Schlick group has proposed the idea of "non-RNA-like" RNA topologies in their construction of an RNA topology database (97,99,100,142). However, no work has yet focused on the construction of decoy RNA structures. In this work, we propose a novel approach for generating decoy RNA structures based on the XIOS graphical framework. Multiple functional RNA families have been explored, and our decoy RNA structures preserve most of the properties of different functional families.

5.2 Methods

5.2.1 Natural motif database construction

A list of 206 curated RNA structures from 8 functional families have been collected from a variety of resources (see Table 3.3). These RNA structures have been converted to XIOS graphs (Figure 2.1). The fingerprint of each, which is the set of subgraphs of 3 to 7 vertices for each RNA structure, is computed using a subgraph random sampling algorithm (Chapter 3). Each non-isomorphic subgraph has a unique index in a pre-computed RNA structural motif database containing 55,728 motifs (Table 2.1). In addition, the subgraphs of subgraphs have been identified and stored in the motif database (Figure 3.1). The set of subgraphs plus their parental subgraphs are called the extended fingerprint of a RNA graph. Each RNA fingerprint can be written as a binary vector of 55,728 entries indicating whether each motif exists in the database. For each RNA family, the incidence of each motif is calculated as the number of RNAs in that family that contain the motif, and the probability of the motif is its incidence divided by the total number of RNAs in that family. The total probability of a motif is also calculated across all the families. The collection of all 3 to 7 vertex motifs that are found in a set of curated RNA structures is called the decoy database (Table 5.1), which will be used to construct decoy structures.

5.2.2 Construction of random graphs

Random graphs have been constructed as a control for the decoy graphs. The construction starts from one dome, and is incremented by adding one dome at a time. A new dome is added by randomly picking two intervals (an interval is the space separated by

the two ends of the domes and the sequence) in the existing dome plot. This process proceeds until the random graph reaches the desired number of vertices (stems) (Figure 5.1).

5.2.3 Construction of decoy graphs: family-specific and non-family-specific

Decoy graphs are constructed by an exhaustive enumeration approach. A decoy begins with one dome. The structure is iteratively incremented by adding one dome at a time to the structure. All possible locations of adding the new dome are examined. For each candidate location, its structural fingerprint is calculated using random sampling approach, the fingerprint similarity to natural RNAs computed, and weighted by either the family specific appearance probability (family-specific decoys) or the total appearance probability (non-family-specific decoys) pre-calculated in section 2.1. In each iteration, one candidate is sampled, probability weighted by its fingerprint similarity to natural RNAs, and kept for the next iteration. This process proceeds until the decoy graph reaches the desired number of vertices (stems) (Figure 5.2).

5.2.4 Evaluation

The constructed decoy graphs are evaluated by six metrics: the fraction of “O” edges, the fraction of “I” edges, the ratio of “O” edge number to “I” edge number (called O/I ratio), connectivity, degree centrality, global clustering coefficient. An additional graphical metric is the topological fingerprint. Each graph metric is described below. For each RNA functional family, 10 decoy graphs (using motifs from that specific family) and 10

control graphs (using motifs from all RNA families) are generated respectively, with the size (stem number) as the average number of stems in that family. A family-specific decoy graph, a non-family-specific decoy graph, or a random graph is evaluated by its error, which is considered to be the sum of the difference in the six graph metrics, between this graph and curated graphs in a certain family, and also evaluated by the topological fingerprint similarity between this graph and curated graphs in a family.

Fraction of "O" edges

In an RNA structure, the fraction of "O" edges reflects the proportion of pseudoknots, which is one of the most important structural elements in RNA (112).

Fraction of "I" edges

In an RNA structure, the fraction of "I" edges reflects the proportion of nested stems (also called embedding), one of the structural principles in RNA (94,96). For example, the classical structure of the Iron Response Element (IRE), is a hairpin embedded inside of a stem (166,167).

O/I ratio

The O/I ratio reflects the ratio of pseudoknots to that of nested stems, which reveals whether the structure is more pseudoknot-concentrated or nesting-concentrated.

Connectivity

Connectivity is a graph theoretical concept that calculates the minimum number of nodes or edges (nodes are used here) that need to be removed to produce a disconnected graph (168). Some RNA structures (graphs) are less connected, but more organized than others; they are organized as several modules (motifs), and each module contains several stems (vertices), with some stems being the junction connecting different modules.

Degree centrality

Degree centrality measures the number of neighbors (connections) of a vertex. The degree centrality of a graph measures the average in number of connections among the vertices in the graph (169). The degree centrality of RNA structures differs by family, and could therefore be considered an important metric for RNA structures. The calculation of degree centrality of a graph is shown below, as equation (1), N is the total number of vertices in the graph, $\deg(n^*)$ is the maximum degree among all the nodes in the graph, and $\deg(i)$ is the degree of node i:

$$C_D = \frac{\sum_{i=1}^{N}[deg(n^*) - \deg(i)]}{[(N-1)(N-2)]} \quad (1)$$

Global clustering coefficient

The clustering coefficient measures whether the neighbors of a node tend to cluster together. The global clustering coefficient measures the global property of a graph being a

community (170). RNA structures in different families differ in their clustering coefficient, and therefore, global clustering coefficient can be used as a metric in graph comparison. The calculation of global clustering coefficient is shown below, as equation (2):

$$GCC = \frac{number\ of\ closed\ triplets}{number\ of\ connected\ triplets\ of\ vertices} \quad (2)$$

Number of hairpin loops

A hairpin loop in RNA is defined as a loop enclosed by a stem. The number of hairpin loops is one of the important characteristics that discriminate between RNA families (171).

Number of internal loops

An internal loop in RNA is defined as a loop enclosed by two stems. A typical internal loop is comprised of two unpaired areas, however, a special case exists when the loop only contains one unpaired area, and it is called a bulge loop. In this work, bulge loops and typical internal loops are not discriminated. Just like the number of hairpin loops, the number of internal loops is another important characteristic to discriminate RNA families.

Number of multi-loops

A multi-loop in RNA is defined as a stem enclosed by two or more stems. Multi-loops are characteristic of some RNA families, for example, the clover-leaf structure in tRNA is a four-way multi-loop. Therefore, the number of multi-loops is another important RNA structural statistic.

Depth of stem nesting

One stem is considered to be nested in another stem when this stem lies in the loop region of the other stem. Some RNAs have multiple layers of stem nesting, and the depth of nesting is an important RNA structural characteristic.

Structural fingerprint similarity

The structural fingerprints of RNAs R_X and R_Y are represented as two sets of subgraphs: $X = \{g_{X_1}, g_{X_2}, \dots, g_{X_n}\}$ and $Y = \{g_{Y_1}, g_{Y_2}, \dots, g_{Y_m}\}$, where $g_{X_1}, g_{X_2}, \dots, g_{X_n}$ and $g_{Y_1}, g_{Y_2}, \dots, g_{Y_m}$ are subgraphs in the RNAs R_X and R_Y, respectively. Their similarity is measured by the Jaccard Similarity (151), S_J (Chapter 3), as equation (3):

$$S_J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} \tag{3}$$

5.3 Results

The decoy graphs (both family-specific and non-family-specific) are evaluated using the graph features discussed above, using natural RNA graphs as positive controls, and ran-

dom graphs as negative controls. It is expected that the graphs built with family specificity share the most fingerprint similarity with the natural graphs, and the random graphs share the least fingerprint similarity with the natural graphs.

Structural fingerprint similarity

The structural fingerprint similarity to the natural graphs, is highest in the decoy graphs with family specificity, lower in the decoy graphs without family specificity, and lowest in the random graphs (Figure 5.3).

Similarity of decoys to natural graphs

In general, the decoy graphs are more similar to natural graphs than to the random graphs. This is true for degree centrality, fraction of “I” edges, fraction of “O” edges, and number of hairpins. However, the random graphs are more similar to the natural graphs in the depth of nesting. For other features, such as the clustering coefficient, the decoy graphs and the random graphs are equally similar to the natural graphs. In addition, both decoy graphs and random graphs differ from natural graphs in the number of internal/bulge loops, and the number of multi-loops (see Figures 5.4 - 5.13 for more details).

5.4 Discussions

Decoys are synthetic graphs that have some structural similarity to natural RNA graphs. In this work, we have constructed decoy graphs and tested their difference from natural

RNA graphs using several graph features and topological fingerprint similarity. We have shown that decoys are different from random graphs, and they serve as better biological controls in experiments when compared with random graphs, as they have higher fingerprint similarity to the natural graphs, and they are more similar to natural graphs in metrics such as degree centrality, fraction of “I” edges, fraction of “O” edges, and number of hairpins. This is the first work so far to create decoy structures for use in RNA structural studies.

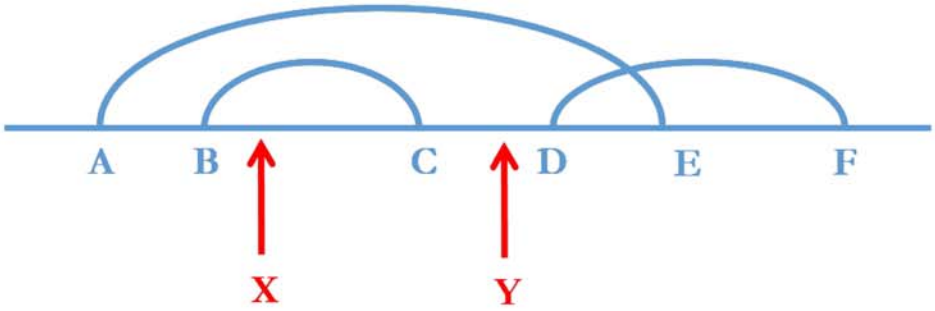


Figure 5.1 Construction of random graphs. Randomly pick two intervals from {(-∞, A), (A,B), (B,C), (C,D), (D,E), (E,F), (F, +∞)}, and these two intervals can be the same. Insert X and Y, the left and right sides of stem (X, Y), into the two intervals, separately.

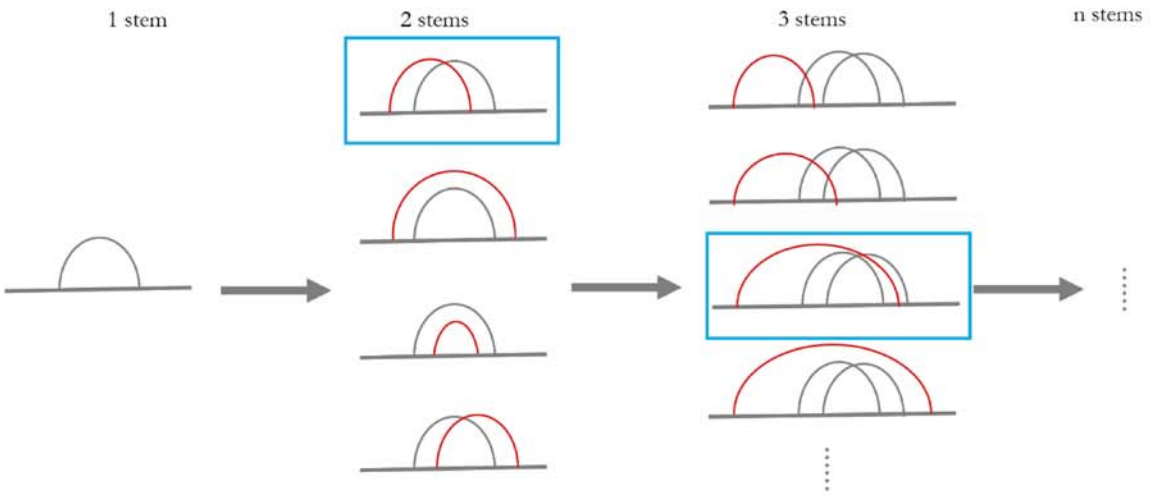


Figure 5.2 Construction of decoy graphs. The construction starts from one stem. In each iteration, one stem is added. When adding the new stem, all the possible combinations are considered, evaluated by their fingerprint similarity to the curated graphs, and one decoy graph is sampled by probability proportional to the fingerprint similarity and the construction of the growing decoy is based on the sampled graph in the next iteration. The construction of decoy graph continues until the number of vertices has reached the desired number.

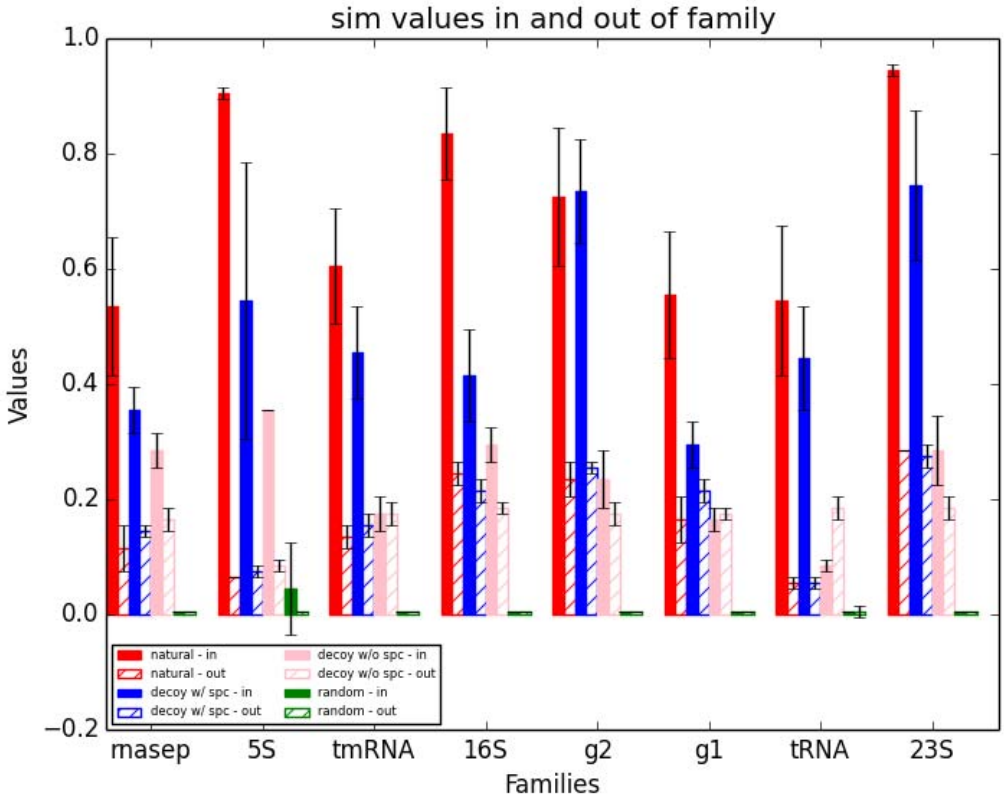


Figure 5.3 Structural fingerprint similarity between: natural RNA graphs (red), decoy graphs with family specificity (blue), decoy graphs without family specificity (pink), or random graphs (green), and natural RNA graphs in the corresponding family (solid) or outside of that family (tilted lines).

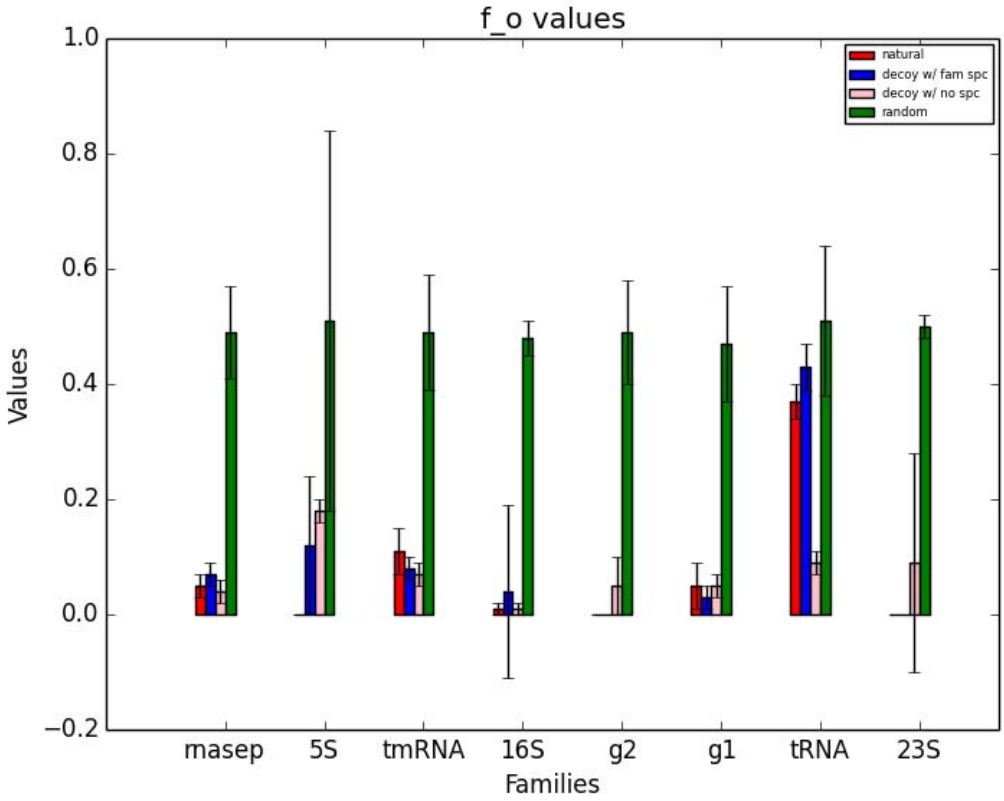


Figure 5.4 Fraction of "O" edges in: natural RNA graphs (red), decoy graphs with family specificity (blue), decoy graphs without family specificity (pink), or random graphs (green).

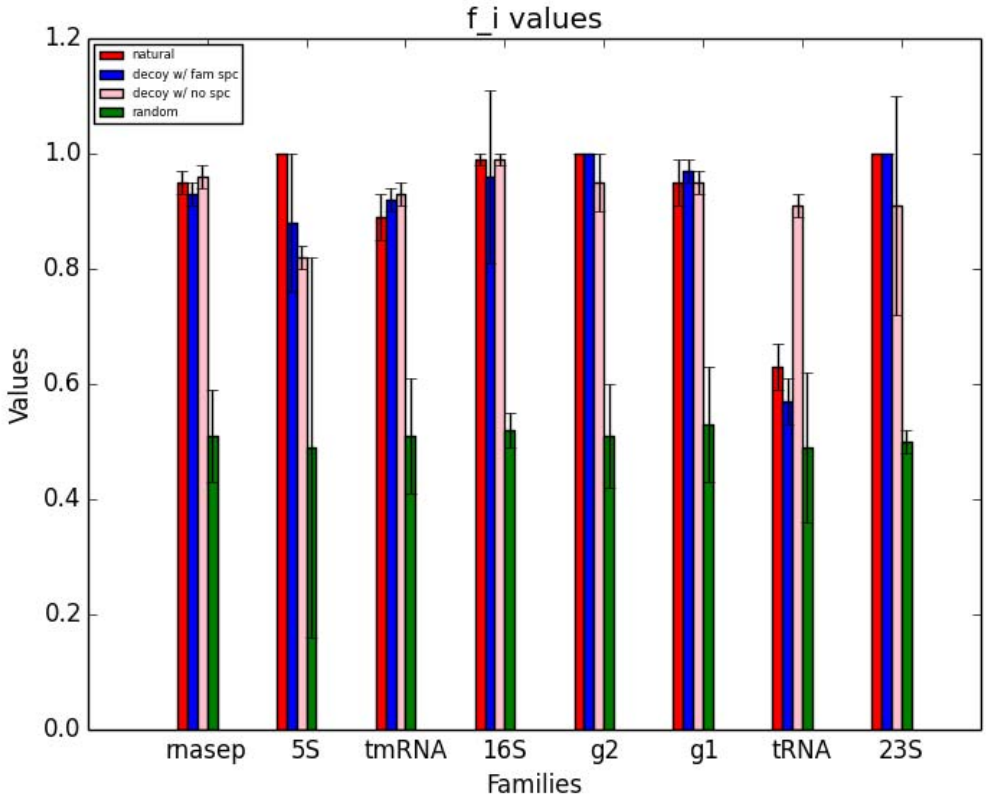


Figure 5.5 Fraction of "I" edges in: natural RNA graphs (red), decoy graphs with family specificity (blue), decoy graphs without family specificity (pink), or random graphs (green).

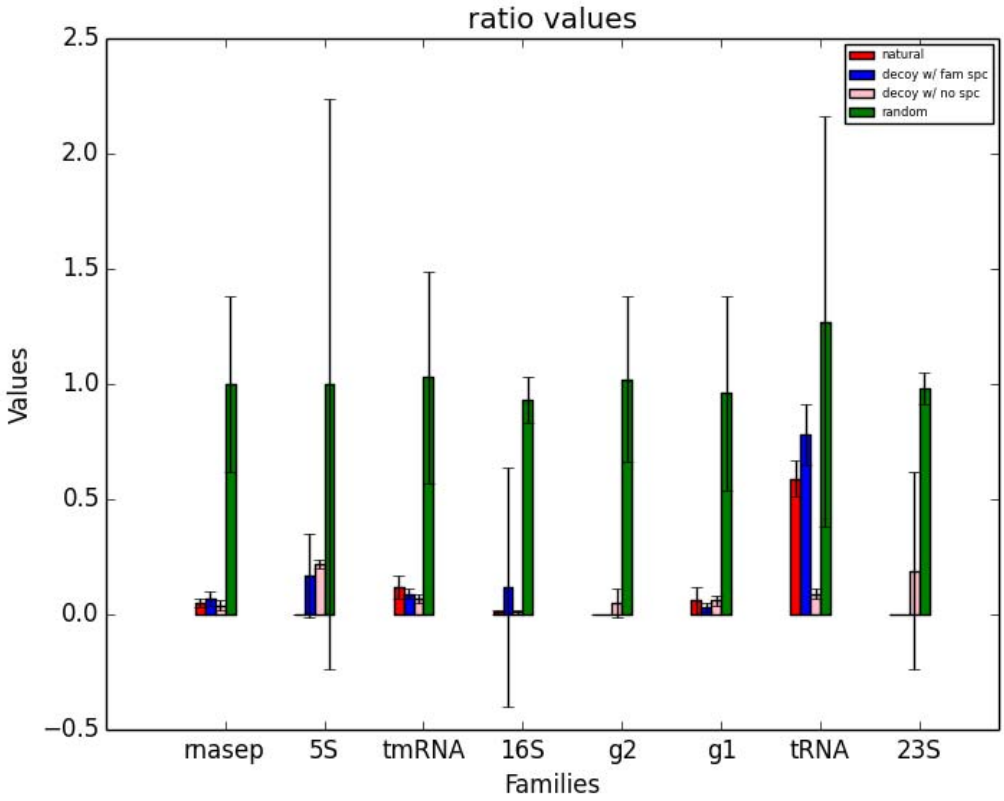


Figure 5.6 Ratio of “O” to “I” values in: natural RNA graphs (red), decoy graphs with family specificity (blue), decoy graphs without family specificity (pink), or random graphs (green).

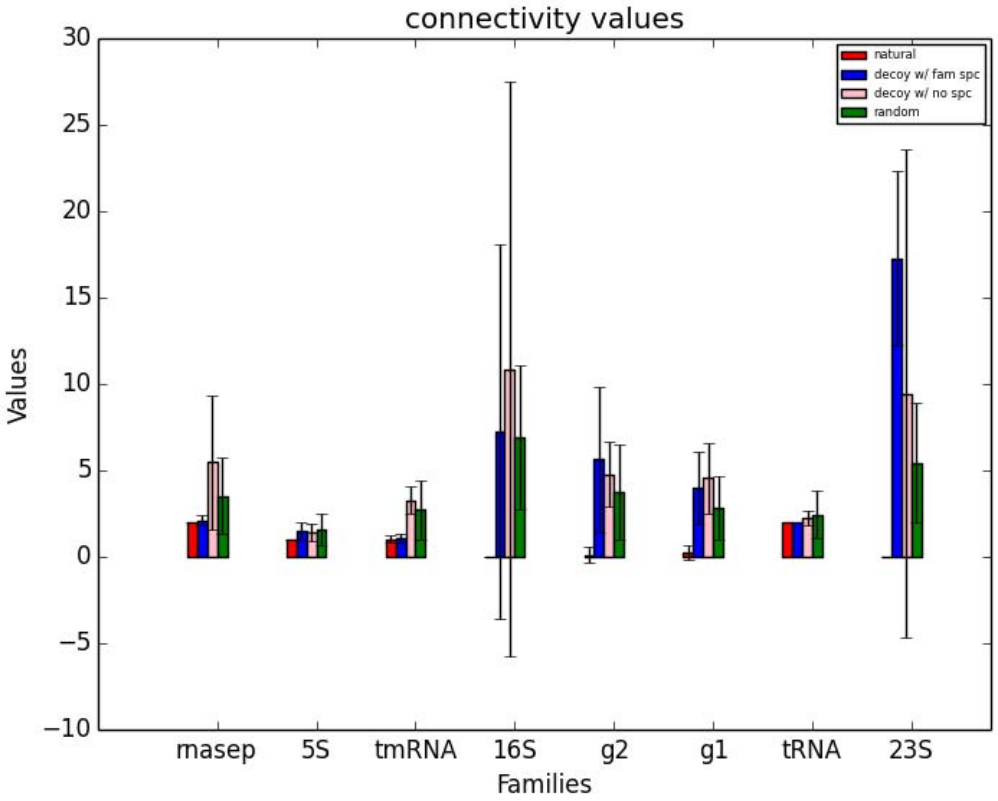


Figure 5.7 Connectivity values in: natural RNA graphs (red), decoy graphs with family specificity (blue), decoy graphs without family specificity (pink), or random graphs (green).

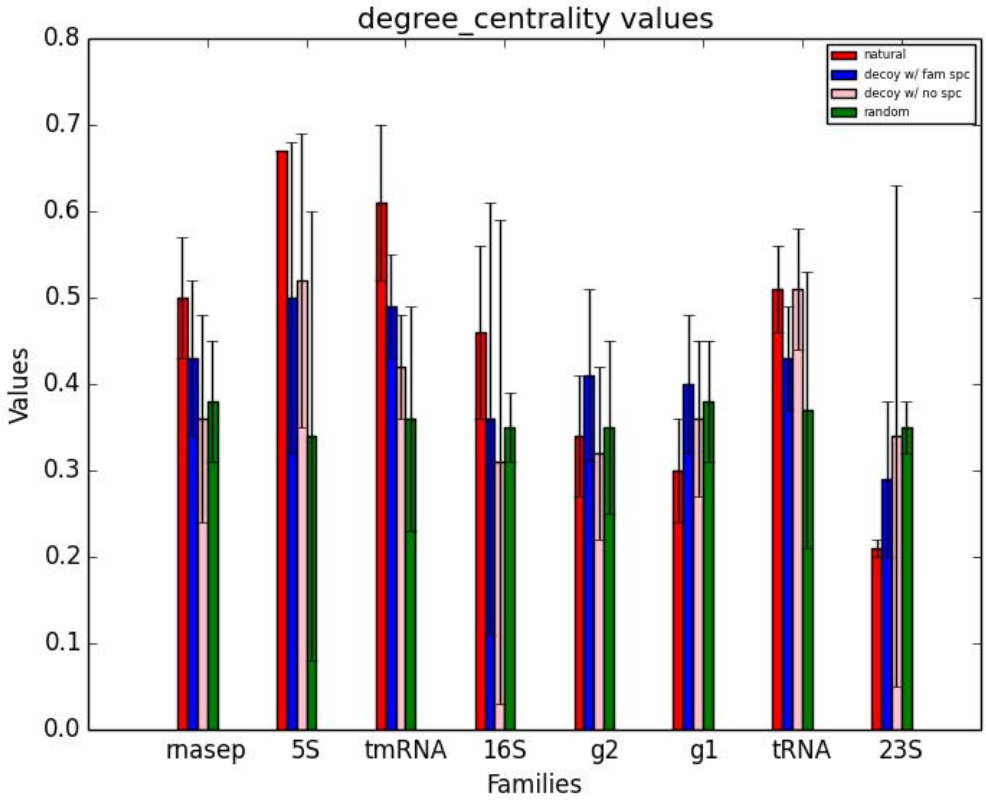


Figure 5.8 Degree centrality values in: natural RNA graphs (red), decoy graphs with family specificity (blue), decoy graphs without family specificity (pink), or random graphs (green).

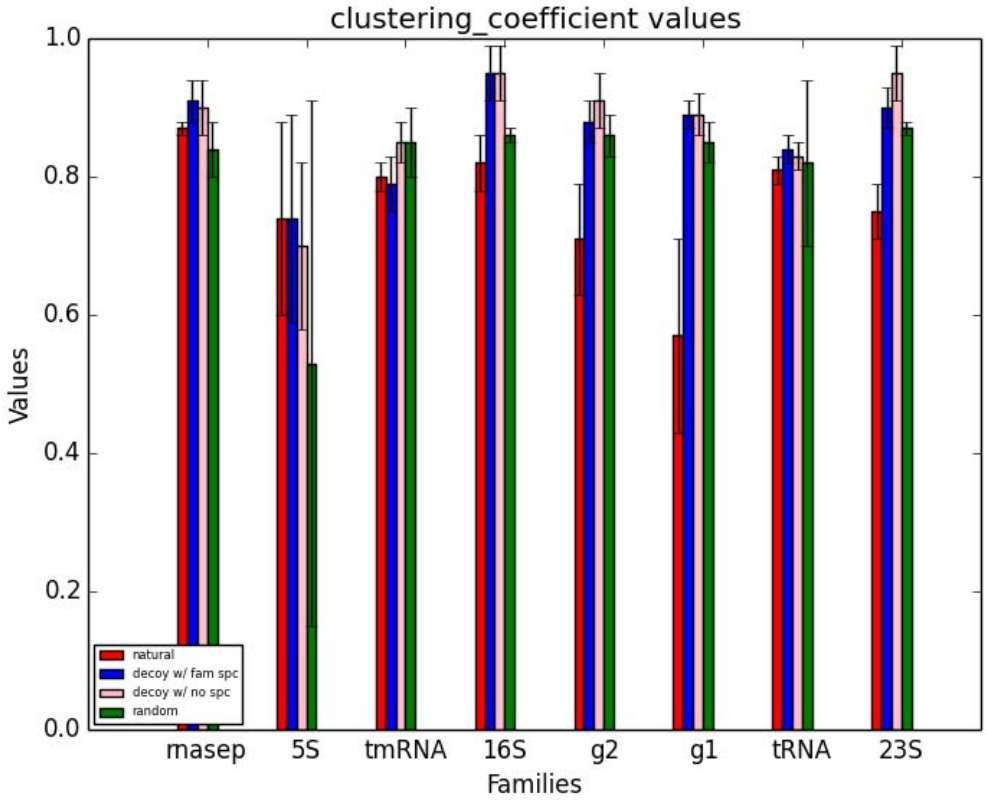


Figure 5.9 Global clustering coefficient values in: natural RNA graphs (red), decoy graphs with family specificity (blue), decoy graphs without family specificity (pink), or random graphs (green).

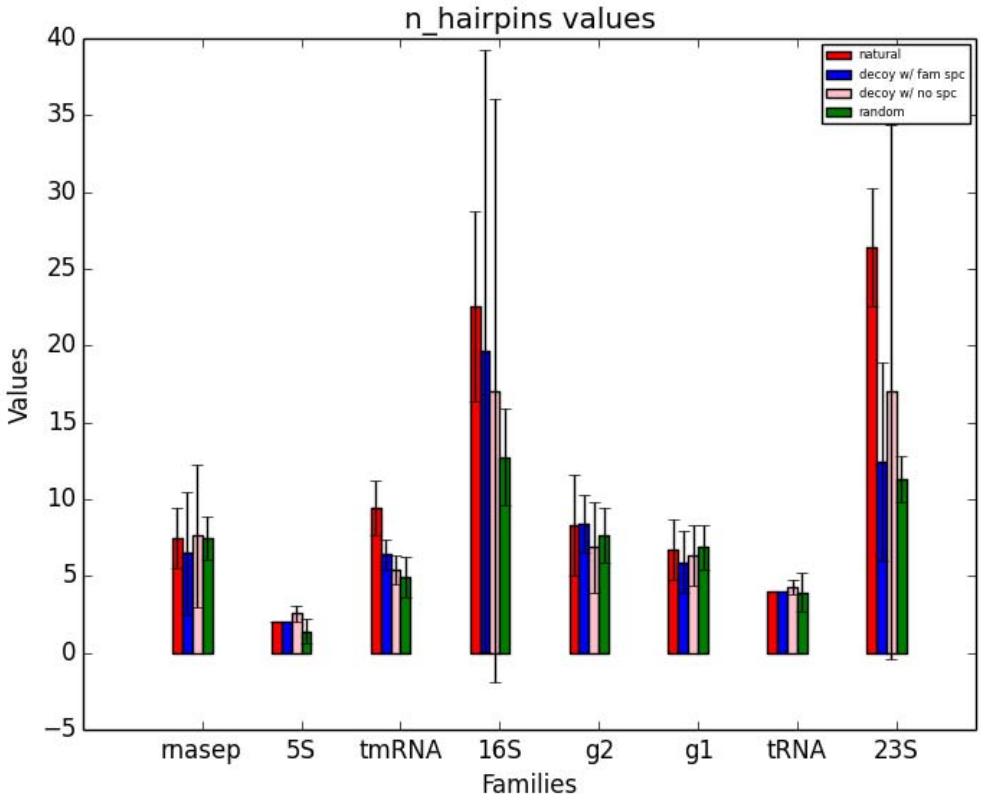


Figure 5.10 Number of hairpin loops in: natural RNA graphs (red), decoy graphs with family specificity (blue), decoy graphs without family specificity (pink), or random graphs (green).

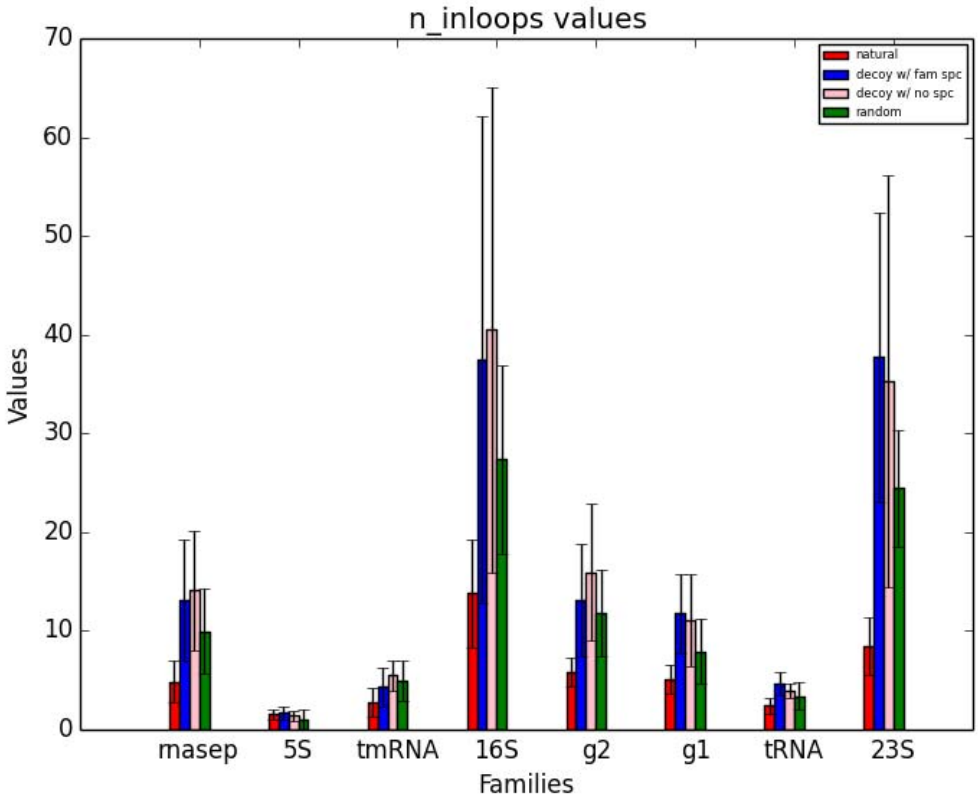


Figure 5.11 Number of internal/bulge loops in: natural RNA graphs (red), decoy graphs with family specificity (blue), decoy graphs without family specificity (pink), or random graphs (green).

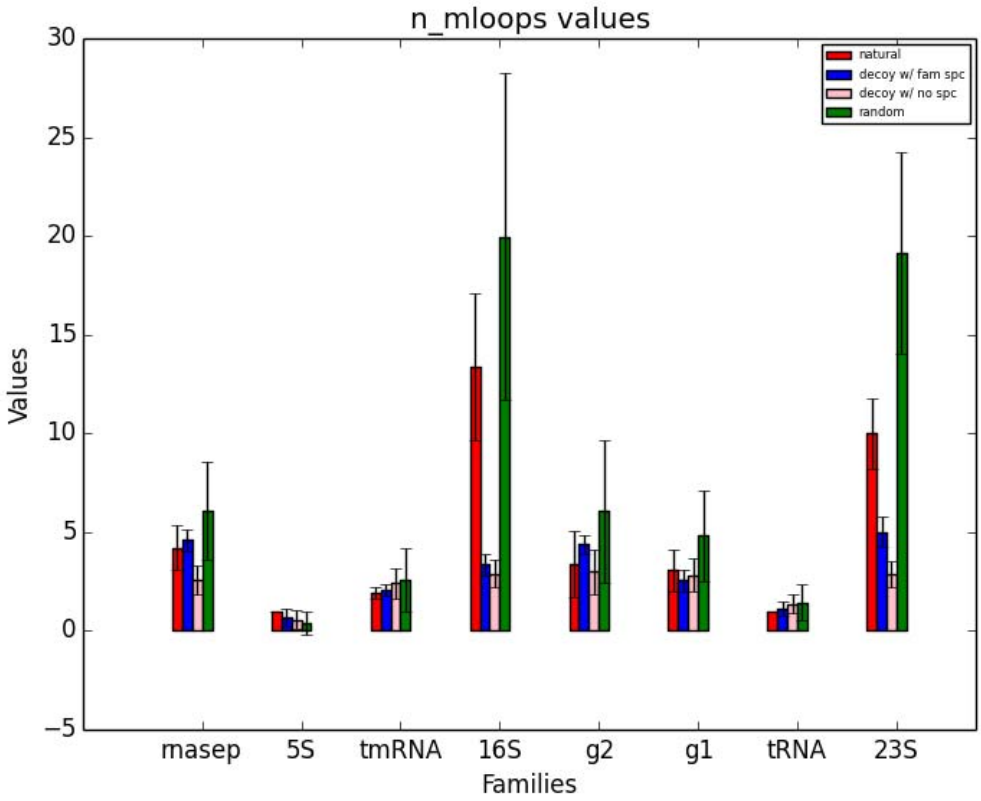


Figure 5.12 Number of multi-loops in: natural RNA graphs (red), decoy graphs with family specificity (blue), decoy graphs without family specificity (pink), or random graphs (green).

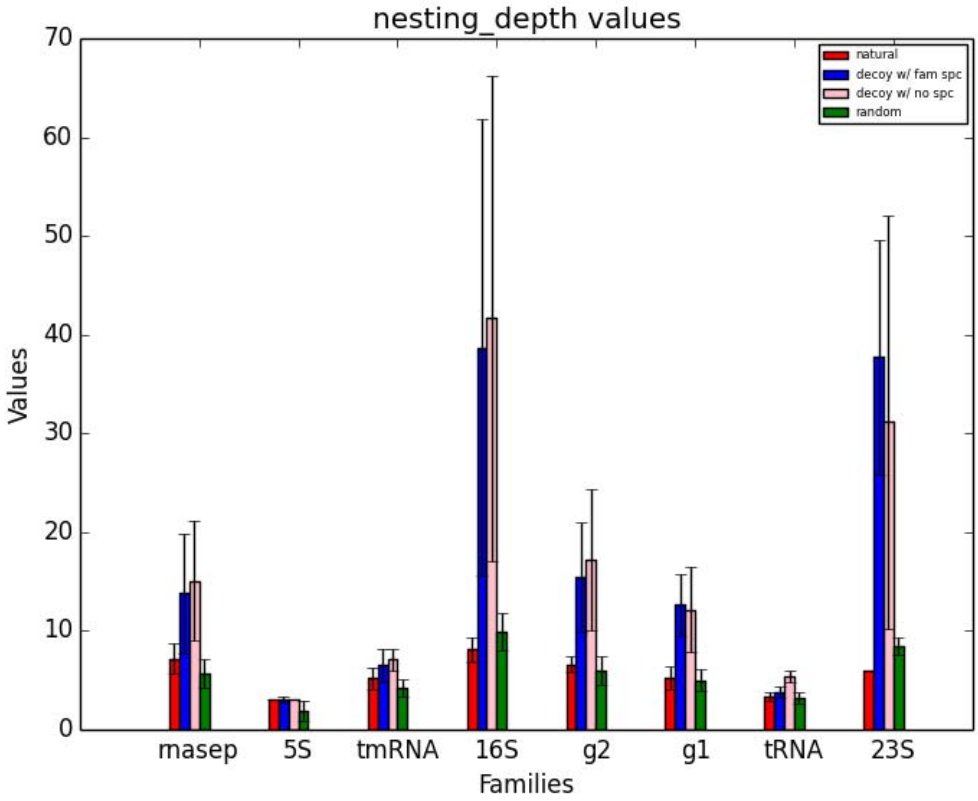


Figure 5.13 Number of stem nesting in: natural RNA graphs (red), decoy graphs with family specificity (blue), decoy graphs without family specificity (pink), or random graphs (green).

Table 5.1 Natural motif database.

Motif size	Naturally existing motif number	Physically possible motif number
3	8	8
4	43	46
5	187	368
6	549	3,914
7	1,241	51,390
Total	2,028	55,726

CHAPTER 6. SUMMARY AND FUTURE DIRECTIONS

6.1 Summary: RNA sequence, topology, and function

As one of the three major macromolecules (DNA, RNA, and protein) in all known forms of life, RNA (ribonucleic acid) is a ubiquitous molecule containing different classes, and that performs multiple cellular functions including regulation and expression of genes. Only some RNA functions are well studied, for example, RNA as a genetic information carrier from DNA to protein (as it was originally known). However, as modern technologies thrive (e.g., high-throughput sequencing), a plethora of novel RNAs have been found, and their complex functions are largely unknown.

Classical methods for identifying the function of a novel DNA or protein molecule function rely on sequence similarity to annotated molecules; however, functional RNA molecules lack a reliable signal at the sequence level. Instead, RNA sequences with similar functions have conserved secondary and higher-order structures; therefore, we can reveal novel RNA functions by structural comparison to known RNA molecules. Moreover, functional motifs, which are the key substructures responsible for RNA function, can be found within groups of RNA molecules with common functions.

In the XIOS graphical framework, RNA structures are represented as graphs; common substructures are identified by finding common (isomorphic) subgraphs among multiple structures. I have developed a subgraph sampling algorithm to efficiently identify sub-

graphs in an RNA graph without exhaustive search. The set of subgraphs in an RNA structure, called its fingerprint, can be compared with the fingerprints of other RNA structures to identify conserved or common structural motifs. I have also developed a distance function for comparing RNA structural fingerprints and demonstrate that it is able of correctly identifying the similarity between structures in known classes of RNA structures. The identification of RNAs with similar structural motifs is a step towards structure-based prediction of RNA function.

Given the importance of RNA structure in understanding its function, it is critical to obtain reliable RNA structures. However, the folded structures of cellular RNAs could be complicated, and it could be difficult and expensive to determine structures using traditional approaches, such as NMR or X-ray crystallography. Computational prediction, cheaper alternative for determining RNA structures, has been applied for 40 years. However, structures predicted by individual programs are only partially correct. I have evaluated multiple combinations of individual programs to find the optimal combination for improved prediction accuracy. Improvement of structure prediction is another step towards the identification of functional RNAs using sequence information. From sequence to structure (topology), and from topology to function, connection of these three dots has provided the foundation for a non-coding RNA BLAST program, through which one can predict the function of a novel RNA based on conserved structural elements.

Based on the XIOS graph framework, I have developed an approach to create decoy RNA structures that serve as better experimental controls than random structures. Decoy

structures are non-biological structures generated by computers, with similar topological characteristics to natural molecules, but having no real biological functions. Although decoy structures are commonly used in protein structure analysis, the work proposed here for generating RNA decoys is the first method for generating RNA topological decoys. Figure 6.1 shows the roadmap of the three major works that make the core of this thesis.

6.2 Future direction: motif distributions in RNA structures

Motifs in a fingerprint are not equally important in RNA classification. One might assume that RNA structural motifs would play a similar role to words in identifying documents with similar content. That is that neither the most frequent nor the least frequent words are the most informative. In information retrieval (172), this is measured by the term frequency (TF) and Inverse Document Frequency (IDF). RNA motifs that are contained only in specific families of RNAs are more important in classification than the ones that appear in most of the RNAs, and motifs that occur in only one or a few structures are similarly uninformative. Using the fingerprints of our curated data (Chapter 3), the curve of the natural log of frequency versus the natural log of rank shows that the distribution of RNA motifs follows Zipf's Law (Figure 6.2).

We use Inverse Document Frequency (IDF) to emphasize the effects on motifs with lower support. Suppose R is the total set of RNA structures (corresponding to documents), r is any one of the structures, and m is a motif (corresponding to a term), then the IDF of motif m in RNA r is

$$IDF_{m;r} = 1 - \frac{|\{r \in R: m \in r\}|}{|R|} \tag{4}$$

IDF is offset by Term Frequency (TF), the frequency of motifs in each RNA structure, which is currently only a Boolean relationship

$$TF_{m;r} = \begin{cases} 1, & m \in r \\ 0, & m \notin r \end{cases} \tag{5}$$

And Term Frequency-Inverse Document Frequency (TF-IDF) is the product of TF and IDF

$$TFIDF_{m;r} = \begin{cases} 1 - \frac{|\{r \in R: m \in r\}|}{|R|}, & m \in r \\ 0, & m \notin r \end{cases} \tag{6}$$

We applied TF-IDF weighting on the Cosine Similarity calculation (173) and achieved better classification results for both simple fingerprints and extended fingerprints than the results using no TF-IDF weighting (Figure 6.3). Other options for measuring term frequency, such as logarithmically scaled frequency, or augmented frequency, could be explored aiming for further improvement in the classification in future experiments.

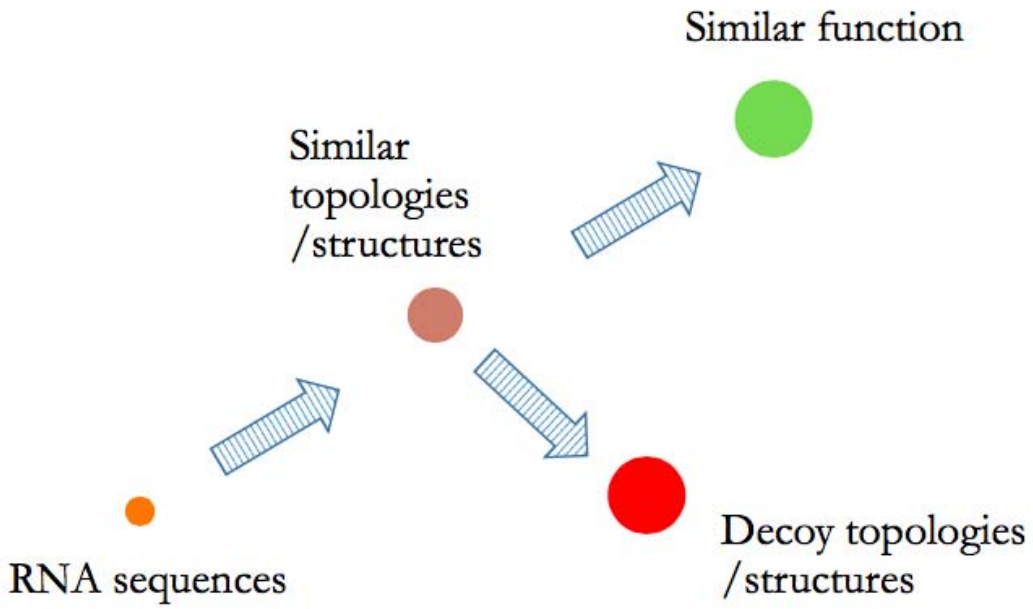


Figure 6.1 Roadmap of the three works: RNA fingerprint, RNA structure prediction improvement, and decoy structure generation.

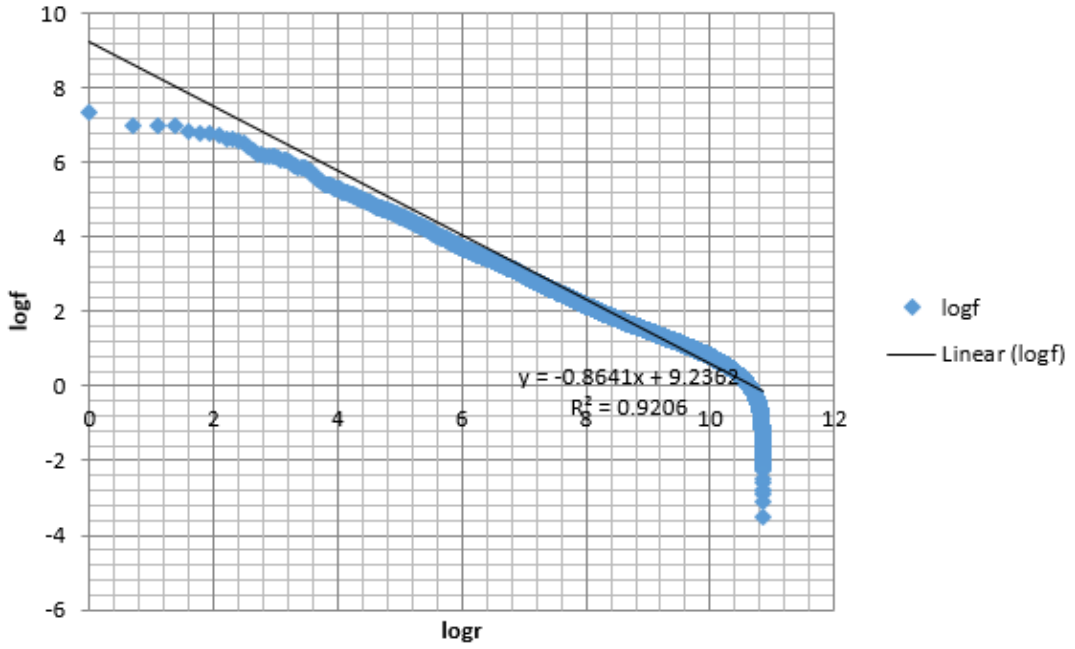


Figure 6.2 RNA motif frequency v.s. rank. This curve is close to a straight line, which indicates that the distribution of RNA motifs approximates Zipf's Law.

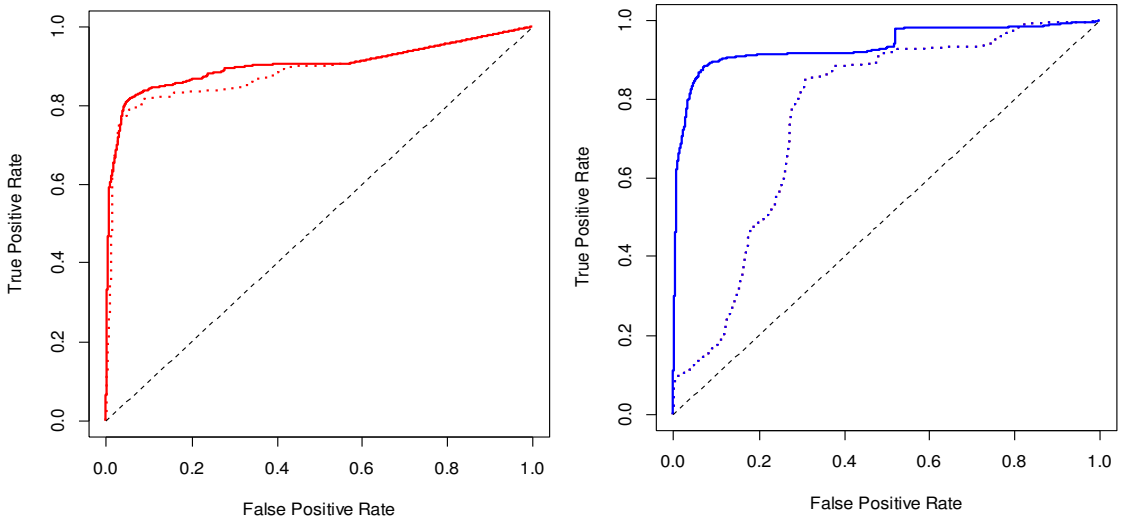


Figure 6.3 TFIDF Weighting in RNA using Cosine Similarity. Left: AUC of SimFP (no weighting, dashed line, 0.802; TFIDF weighting, solid line, 0.902); right: AUC of ExtFP (0.796; 0.936).

LIST OF REFERENCES

LIST OF REFERENCES

1. Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G. *et al.* (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science*, **308**, 1149-1154.
2. Plank, T.D. and Kieft, J.S. (2012) The structures of nonprotein-coding RNAs that drive internal ribosome entry site function. *Wiley Interdiscip Rev RNA*, **3**, 195-212.
3. Zuker, M., Mathews, D.H. and Turner, D.H. (1999) Algorithms and thermodynamics for RNA secondary structure prediction: A practical guide. *Rna Biochemistry and Biotechnology*, **70**, 11-43.
4. Hermann, T. and Westhof, E. (1999) Non-Watson-Crick base pairs in RNA-protein recognition. *Chem Biol*, **6**, R335-343.
5. Gultyaev, A.P., Olsthoorn, R. C., Pleij, C. W. and Westhof, E. (2012) *RNA Structure: Pseudoknots*. John Wiley and Sons, Inc., eLS.
6. Staple, D.W. and Butcher, S.E. (2005) Pseudoknots: RNA structures with diverse functions. *Plos Biology*, **3**, 956-959.
7. Altman, S. (2001) *The RNA world*, http://www.nobelprize.org/nobel_prizes/chemistry/laureates/1989/altman-article.html.
8. Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N. and Altman, S. (1983) The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, **35**, 849-857.
9. Cech, T. (2004) *Exploring the New RNA World,* http://www.nobelprize.org/nobel_prizes/chemistry/laureates/1989/cech-article.html.
10. Kruger, K., Grabowski, P.J., Zaug, A.J., Sands, J., Gottschling, D.E. and Cech, T.R. (1982) Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell*, **31**, 147-157.
11. Robertson, M.P. and Joyce, G.F. (2012) The origins of the RNA world. *Cold Spring Harb Perspect Biol*, **4**.
12. Gilbert, W. (1986) The RNA World. *Nature*, **319**, 618.
13. Esteller, M. (2011) Non-coding RNAs in human disease. *Nat Rev Genet*, **12**, 861-874.
14. Feng, J., Funk, W.D., Wang, S.S., Weinrich, S.L., Avilion, A.A., Chiu, C.P., Adams, R.R., Chang, E., Allsopp, R.C. and Yu, J. (1995) The RNA component of human telomerase. *Science*, **269**, 1236-1241.

15. Lehmann, K. and Schmidt, U. (2003) Group II introns: structure and catalytic versatility of large natural ribozymes. *Crit Rev Biochem Mol Biol*, **38**, 249-303.
16. Paquin, B.a.S., D. A. (2001) *Introns: Group I Structure and Function.* . John Wiley & Sons, Ltd. , eLS.
17. Naville, M. and Gautheret, D. (2010) Transcription attenuation in bacteria: theme and variations. *Brief Funct Genomics*, **9**, 178-189.
18. Vitreschak, A.G., Rodionov, D.A., Mironov, A.A. and Gelfand, M.S. (2004) Riboswitches: the oldest mechanism for the regulation of gene expression? *Trends Genet*, **20**, 44-50.
19. Bompfunewerer, A.F., Flamm, C., Fried, C., Fritzsch, G., Hofacker, I.L., Lehmann, J., Missal, K., Mosig, A., Muller, B., Prohaska, S.J. *et al.* (2005) Evolutionary patterns of non-coding RNAs. *Theory in Biosciences*, **123**, 301-369.
20. Wower, I.K., Zwieb, C. and Wower, J. (2005) Transfer-messenger RNA unfolds as it transits the ribosome. *RNA*, **11**, 668-673.
21. Pavesi, G., Mauri, G., Stefani, M. and Pesole, G. (2004) RNAProfile: an algorithm for finding conserved secondary structure motifs in unaligned RNA sequences. *Nucleic Acids Research*, **32**, 3258-3269.
22. Torarinsson, E., Sawera, M., Havgaard, J.H., Fredholm, M. and Gorodkin, J. (2006) Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. *Genome Research*, **16**, 885-889.
23. Bessho, Y., Shibata, R., Sekine, S., Murayama, K., Higashijima, K., Hori-Takemoto, C., Shirouzu, M., Kuramitsu, S. and Yokoyama, S. (2007) Structural basis for functional mimicry of long-variable-arm tRNA by transfer-messenger RNA. *Proc Natl Acad Sci U S A*, **104**, 8293-8298.
24. Weis, F., Bron, P., Rolland, J.P., Thomas, D., Felden, B. and Gillet, R. (2010) Accommodation of tmRNA-SmpB into stalled ribosomes: a cryo-EM study. *RNA*, **16**, 299-306.
25. Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S. *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res*, **22**, 1760-1774.
26. Pace, N.R. and Brown, J.W. (1995) Evolutionary perspective on the structure and function of ribonuclease P, a ribozyme. *J Bacteriol*, **177**, 1919-1928.
27. Ellis, J.C. and Brown, J.W. (2009) The RNase P family. *RNA Biology*, **6**, 362-369.
28. Hedberg, A. and Johansen, S.D. (2013) Nuclear group I introns in self-splicing and beyond. *Mob DNA*, **4**, 17.
29. Adams, P.L., Stahley, M.R., Kosek, A.B., Wang, J. and Strobel, S.A. (2004) Crystal structure of a self-splicing group I intron with both exons. *Nature*, **430**, 45-50.
30. Ke, A., Ding, F., Batchelor, J.D. and Doudna, J.A. (2007) Structural roles of monovalent cations in the HDV ribozyme. *Structure*, **15**, 281-287.
31. Tijerina, P., Mohr, S. and Russell, R. (2007) DMS footprinting of structured RNAs and RNA-protein complexes. *Nat Protoc*, **2**, 2608-2623.

32. Fritz, J.J., Lewin, A., Hauswirth, W., Agarwal, A., Grant, M. and Shaw, L. (2002) Development of hammerhead ribozymes to modulate endogenous gene expression for functional studies. *Methods*, **28**, 276-285.
33. Shaw, L.C. and Lewin, A.S. (1995) Protein-induced folding of a group I intron in cytochrome b pre-mRNA. *J Biol Chem*, **270**, 21552-21562.
34. Noller, H.F. and Chaires, J.B. (1972) Functional modification of 16S ribosomal RNA by kethoxal. *Proc Natl Acad Sci U S A*, **69**, 3115-3118.
35. Ziehler, W.A. and Engelke, D.R. (2001) Probing RNA structure with chemical reagents and enzymes. *Curr Protoc Nucleic Acid Chem*, **Chapter 6**, Unit 6.1.
36. Mathews, D.H., Disney, M.D., Childs, J.L., Schroeder, S.J., Zuker, M. and Turner, D.H. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci U S A*, **101**, 7287-7292.
37. Merino, E.J., Wilkinson, K.A., Coughlan, J.L. and Weeks, K.M. (2005) RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). *J Am Chem Soc*, **127**, 4223-4231.
38. Mortimer, S.A. and Weeks, K.M. (2007) A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. *J Am Chem Soc*, **129**, 4144-4145.
39. Deigan, K.E., Li, T.W., Mathews, D.H. and Weeks, K.M. (2009) Accurate SHAPE-directed RNA structure determination. *Proc Natl Acad Sci U S A*, **106**, 97-102.
40. Knapp, G. (1989) Enzymatic approaches to probing of RNA secondary and tertiary structure. *Methods Enzymol*, **180**, 192-212.
41. Kubota, M., Tran, C. and Spitale, R.C. (2015) Progress and challenges for chemical probing of RNA structure inside living cells. *Nat Chem Biol*, **11**, 933-941.
42. Kertesz, M., Wan, Y., Mazor, E., Rinn, J.L., Nutter, R.C., Chang, H.Y. and Segal, E. (2010) Genome-wide measurement of RNA secondary structure in yeast. *Nature*, **467**, 103-107.
43. Underwood, J.G., Uzilov, A.V., Katzman, S., Onodera, C.S., Mainzer, J.E., Mathews, D.H., Lowe, T.M., Salama, S.R. and Haussler, D. (2010) FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nature Methods*, **7**, 995-U981.
44. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. and Weissman, J.S. (2014) Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature*, **505**, 701-705.
45. Lucks, J.B., Mortimer, S.A., Trapnell, C., Luo, S., Aviran, S., Schroth, G.P., Pachter, L., Doudna, J.A. and Arkin, A.P. (2011) Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proc Natl Acad Sci U S A*, **108**, 11063-11068.
46. Rivas, E., Lang, R. and Eddy, S.R. (2012) A range of complex probabilistic models for RNA secondary structure prediction that includes the nearest-neighbor model and more. *RNA*, **18**, 193-212.

47. Zuker, M. and Sankoff, D. (1984) RNA secondary structures and their prediction. *Bulletin of Mathematical Biology*, **46**, 591-621.
48. Mathews, D.H., Sabina, J., Zuker, M. and Turner, D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol*, **288**, 911-940.
49. SantaLucia, J. and Turner, D.H. (1997) Measuring the thermodynamics of RNA secondary structure formation. *Biopolymers*, **44**, 309-319.
50. Tinoco, I., Borer, P.N., Dengler, B., Levin, M.D., Uhlenbeck, O.C., Crothers, D.M. and Bralla, J. (1973) Improved estimation of secondary structure in ribonucleic acids. *Nat New Biol*, **246**, 40-41.
51. Borer, P.N., Dengler, B., Tinoco, I. and Uhlenbeck, O.C. (1974) Stability of ribonucleic acid double-stranded helices. *J Mol Biol*, **86**, 843-853.
52. Freier, S.M., Kierzek, R., Jaeger, J.A., Sugimoto, N., Caruthers, M.H., Neilson, T. and Turner, D.H. (1986) Improved free-energy parameters for predictions of RNA duplex stability. *Proc Natl Acad Sci U S A*, **83**, 9373-9377.
53. Xia, T., SantaLucia, J., Burkard, M.E., Kierzek, R., Schroeder, S.J., Jiao, X., Cox, C. and Turner, D.H. (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry*, **37**, 14719-14735.
54. Turner, D.H. and Mathews, D.H. (2010) NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Research*, **38**, D280-282.
55. McCaskill, J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105-1119.
56. Sankoff, D., Kruskal, J.B., Mainville, S. and Cedergren, R.J. (1983) *Fast algorithms to determine RNA secondary structures containing multiple loops*. Addison-Wesley Reading, Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison.
57. Waterman, M.S. and Smith, T.M. (1978) RNA secondary structure: a complete mathematical analysis. *Mathematical Biosciences*, **42**, 257-266.
58. Waterman, M. (ed.) (1978) *Secondary structure of single-stranded nucleic acids*. Academic Press, New York.
59. Nussinov, R., Pieczenik, G., Griggs, J.R. and Kleitman, D.J. (1978) Algorithm for loop matchings. *SIAM J. Appl. Math.*, **35**, 68-82.
60. Nussinov, R. and Jacobson, A.B. (1980) Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc Natl Acad Sci U S A*, **77**, 6309-6313.
61. Zuker, M. and Stiegler, P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, **9**, 133-148.
62. Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, **31**, 3406-3415.

63. Markham, N.R. and Zuker, M. (2008) UNAFold: software for nucleic acid folding and hybridization. *Methods Mol Biol*, **453**, 3-31.
64. Bellaousov, S., Reuter, J.S., Seetin, M.G. and Mathews, D.H. (2013) RNAstructure: Web servers for RNA secondary structure prediction and analysis. *Nucleic Acids Research*, **41**, W471-474.
65. Mathews, D.H. (2006) RNA secondary structure analysis using RNAstructure. *Curr Protoc Bioinformatics*, **Chapter 12**, Unit 12.16.
66. Reuter, J.S. and Mathews, D.H. (2010) RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, **11**, 129.
67. Lorenz, R., Bernhart, S.H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F. and Hofacker, I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol Biol*, **6**, 26.
68. Hofacker, I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Research*, **31**, 3429-3431.
69. Hofacker, I.L. (2004) RNA secondary structure analysis using the Vienna RNA package. *Curr Protoc Bioinformatics*, **Chapter 12**, Unit 12.12.
70. Ding, Y. and Lawrence, C.E. (2003) A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Research*, **31**, 7280-7301.
71. Ding, Y., Chan, C.Y. and Lawrence, C.E. (2005) RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA*, **11**, 1157-1166.
72. Do, C.B., Woods, D.A. and Batzoglou, S. (2006) CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, **22**, e90-98.
73. Do, C.B., Foo, C.S. and Batzoglou, S. (2008) A max-margin model for efficient simultaneous alignment and folding of RNA sequences. *Bioinformatics*, **24**, i68-76.
74. Hamada, M., Kiryu, H., Sato, K., Mituyama, T. and Asai, K. (2009) Prediction of RNA secondary structure using generalized centroid estimators. *Bioinformatics*, **25**, 465-473.
75. Sato, K., Kato, Y., Hamada, M., Akutsu, T. and Asai, K. (2011) IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics*, **27**, i85-93.
76. Lyngsø, R.B. and Pedersen, C.N. (2000) RNA pseudoknot prediction in energy-based models. *J Comput Biol*, **7**, 409-427.
77. Beyer, W. (2010), University of Vienna.
78. Bellaousov, S. and Mathews, D.H. (2010) ProbKnot: fast prediction of RNA secondary structure including pseudoknots. *RNA*, **16**, 1870-1880.
79. Cao, S. and Chen, S.J. (2006) Predicting RNA pseudoknot folding thermodynamics. *Nucleic Acids Research*, **34**, 2634-2652.
80. Cao, S. and Chen, S.J. (2009) Predicting structures and stabilities for H-type pseudoknots with interhelix loops. *RNA*, **15**, 696-706.
81. Janssen, S. and Giegerich, R. (2015) The RNA shapes studio. *Bioinformatics*, **31**, 423-425.
82. Corinna Theis, S.J., and Robert Giegerich. (2010) Prediction of RNA Secondary Structure Including Kissing Hairpin Motifs *Algorithms in Bioinformatics*, **6293**, 52-64.

83. Martinsen, L., Johnsen, A., Venanzetti, F. and Bachmann, L. (2010) Phylogenetic footprinting of non-coding RNA: hammerhead ribozyme sequences in a satellite DNA family of Dolichopoda cave crickets (Orthoptera, Rhaphidophoridae). *BMC Evolutionary Biology*, **10**.
84. Zwieb, C., Wower, I. and Wower, J. (1999) Comparative sequence analysis of tmRNA. *Nucleic Acids Research*, **27**, 2063-2071.
85. Eddy, S.R. and Durbin, R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Research*, **22**, 2079-2088.
86. Mount, D.W. (2004). Bioinformatics: Sequence and Genome Analysis. 2nd ed. Cold Spring Harbor Laboratory Press pp. 345-351.
87. Gutell, R.R., Lee, J.C. and Cannone, J.J. (2002) The accuracy of ribosomal RNA comparative structure models. *Curr Opin Struct Biol*, **12**, 301-310.
88. Nussinov, R., Pieczenik, G., Griggs, J.R. and Kleitman, D.J. (1978) Algorithms for Loop Matchings *SIAM Journal on Applied Mathematics*, **35**, 68-82.
89. Hofacker IL, F.W., Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh Chem*, **125**, 167-188.
90. Shapiro, B.A. (1988) An algorithm for comparing multiple RNA secondary structures. *Comput Appl Biosci*, **4**, 387-393.
91. Shapiro, B.A. and Zhang, K.Z. (1990) Comparing multiple RNA secondary structures using tree comparisons. *Comput Appl Biosci*, **6**, 309-318.
92. Margalit, H., Shapiro, B.A., Oppenheim, A.B. and Maizel, J.V. (1989) Detection of common motifs in RNA secondary structures. *Nucleic Acids Research*, **17**, 4829-4845.
93. Fontana, W., Konings, D.A., Stadler, P.F. and Schuster, P. (1993) Statistics of RNA secondary structures. *Biopolymers*, **33**, 1389-1404.
94. Giegerich, R., Voss, B. and Rehmsmeier, M. (2004) Abstract shapes of RNA. *Nucleic Acids Research*, **32**, 4843-4851.
95. Steffen, P., Voss, B., Rehmsmeier, M., Reeder, J. and Giegerich, R. (2006) RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics*, **22**, 500-503.
96. Janssen, S., Reeder, J. and Giegerich, R. (2008) Shape based indexing for faster search of RNA family databases. *BMC Bioinformatics*, **9**, 131.
97. Gan, H.H., Pasquali, S. and Schlick, T. (2003) Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design. *Nucleic Acids Research*, **31**, 2926-2943.
98. Laing, C. and Schlick, T. (2011) Computational approaches to RNA structure prediction, analysis, and design. *Curr Opin Struct Biol*, **21**, 306-318.
99. Gan, H.H., Fera, D., Zorn, J., Shiffeldrim, N., Tang, M., Laserson, U., Kim, N. and Schlick, T. (2004) RAG: RNA-As-Graphs database--concepts, analysis, and features. *Bioinformatics*, **20**, 1285-1291.
100. Fera, D., Kim, N., Shiffeldrim, N., Zorn, J., Laserson, U., Gan, H.H. and Schlick, T. (2004) RAG: RNA-As-Graphs web resource. *BMC Bioinformatics*, **5**, 88.

101. Byun, Y. and Han, K. (2009) PseudoViewer3: generating planar drawings of large-scale RNA structures with pseudoknots. *Bioinformatics*, **25**, 1435-1437.
102. Silvian, L.F., Wang, J. and Steitz, T.A. (1999) Insights into editing from an ile-tRNA synthetase structure with tRNAile and mupirocin. *Science*, **285**, 1074-1077.
103. Tzanetakis, I.E. and Martin, R.R. (2007) Strawberry chlorotic fleck: identification and characterization of a novel Closterovirus associated with the disease. *Virus Res*, **124**, 88-94.
104. Li, K., Rahman, R., Gupta, A., Siddavatam, P. and Gribskov, M. (2008) Pattern matching in RNA structures. *Bioinformatics Research and Applications*, **4983**, 317-330.
105. Edwards, A.L. and Batey, R.T. (2010) Riboswitches: A Common RNA Regulatory Element. *Nature Education* **3**, 9.
106. Gupta, A., Rahman, R., Li, K. and Gribskov, M. (2012) Identifying complete RNA structural ensembles including pseudoknots. *RNA Biol*, **9**, 187-199.
107. Yan, X.F. and Han, J.W. (2002) gSpan: Graph-based substructure pattern mining. *Proceedings of the 2002 IEEE International Conference on Data Mining*, 721-724.
108. Barrandon, C., Spiluttini, B. and Bensaude, O. (2008) Non-coding RNAs regulating the transcriptional machinery. *Biol Cell*, **100**, 83-95.
109. Pang, K., Frith, M. and Mattick, J. (2006) Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet*, **22**, 1-5.
110. Johnsson, P., Lipovich, L., Grandér, D. and Morris, K. (2014) Evolutionary conservation of long non-coding RNAs; sequence, structure, function. *Biochim Biophys Acta*, **1840**, 1063-1071.
111. Ellis, J.C. and Brown, J.W. (2009) The RNase P family. *RNA Biol*, **6**, 362-369.
112. Staple, D.W. and Butcher, S.E. (2005) Pseudoknots: RNA structures with diverse functions. *PLoS Biol*, **3**, 956-959.
113. Powers, T. and Noller, H.F. (1991) A functional pseudoknot in 16S ribosomal RNA. *EMBO J*, **10**, 2203-2214.
114. Egli, M., Sarkhel, S., Minasov, G. and Rich, A. (2003) Structure and Function of the Ribosomal Frameshifting Pseudoknot RNA from Beet Western Yellow Virus. *Helvetica Chimica Acta*, **86**, 1709-1727.
115. Waterman, M. (1978) Secondary structure of single-stranded nucleic acids. *Adv Math*, **1**, 167-212 (suppl.).
116. Shu, W., Bo, X., Zheng, Z. and Wang, S. (2008) A novel representation of RNA secondary structure based on element-contact graphs. *BMC Bioinformatics*, **9**, 188.
117. Benedetti, G. and Morosetti, S. (1996) A graph-topological approach to recognition of pattern and similarity in RNA secondary structures. *Biophys Chem*, **59**, 179-184.
118. Heyne, S., Costa, F., Rose, D. and Backofen, R. (2012) GraphClust: alignment-free structural clustering of local RNA secondary structures. *Bioinformatics*, **28**, i224-i232.

119. Costa, F. and Grave, K.D. (2010), *Proceedings of the 27th International Conference on Machine Learning (ICML-10)* Omnipress, Haifa, Israel, pp. 255-262.
120. Barash, D. and Comaniciu, D. (2003) A Common Viewpoint on Broad Kernel Filtering and Nonlinear Diffusion. *Lect Notes Comput Sc*, **2695** 683-698.
121. Barash, D. (2004) Spectral Decomposition for the Search and Analysis of RNA Secondary Structure. *J Comput Biol*, **11**, 1169-1174.
122. Rose, P.W., Beran, B., Bi, C., Bluhm, W.F., Dimitropoulos, D., Goodsell, D.S., Prlic, A., Quesada, M., Quinn, G.B., Westbrook, J.D. *et al.* (2011) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Research*, **39**, D392-401.
123. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Research*, **28**, 235-242.
124. Nawrocki, E.P., Burge, S.W., Bateman, A., Daub, J., Eberhardt, R.Y., Eddy, S.R., Floden, E.W., Gardner, P.P., Jones, T.A., Tate, J. *et al.* (2015) Rfam 12.0: updates to the RNA families database. *Nucleic Acids Research*, **43**, D130-137.
125. Andronescu, M., Bereg, V., Hoos, H.H. and Condon, A. (2008) RNA STRAND: the RNA secondary structure and statistical analysis database. *BMC Bioinformatics*, **9**, 340.
126. Petrov, A.S., Bernier, C.R., Hershkovits, E., Xue, Y., Waterbury, C.C., Hsiao, C., Stepanov, V.G., Gaucher, E.A., Grover, M.A., Harvey, S.C. *et al.* (2013) Secondary structure and domain architecture of the 23S and 5S rRNAs. *Nucleic Acids Research*, **41**, 7522-7535.
127. Garey, M.R. and Johnson, D.S. (1979) *Computers and Intractability: A Guide to the Theory of NP-completeness.* W. H. Freeman, New York.
128. Gribskov, M. and Robinson, N.L. (1996) Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Comput Chem*, **20**, 25-33.
129. Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, **4**, 406-425.
130. Reeder, J., Steffen, P. and Giegerich, R. (2007) pknotsRG: RNA pseudoknot folding including near-optimal structures and sliding windows. *Nucleic Acids Research*, **35**, W320-324.
131. Ren, J., Rastegari, B., Condon, A. and Hoos, H.H. (2005) HotKnots: heuristic prediction of RNA secondary structures including pseudoknots. *RNA*, **11**, 1494-1504.
132. Chen, S. and Zhang, K. (2014) An improved algorithm for tree edit distance with applications for RNA secondary structure comparison. *J Comb Optim*, **27**, 778-797.
133. Saito, Y., Sato, K. and Sakakibara, Y. (2011) Fast and accurate clustering of noncoding RNAs using ensembles of sequence alignments and secondary structures. *BMC Bioinformatics*, **12 Suppl 1**, S48.
134. Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. *J Mol Biol*, **147**, 195-197.

135. Kin, T., Tsuda, K. and Asai, K. (2002) Marginalized kernels for RNA sequence data analysis. *Genome Inform*, **13**, 112-122.
136. Karklin, Y., Meraz, R.F. and Holbrook, S.R. (2005) Classification of non-coding RNA using graph representations of secondary structure. *Pac Symp Biocomput*, 4-15.
137. Liu, Q., Zhang, Y., Xu, Y. and Ye, X. (2008) Fuzzy kernel clustering of RNA secondary structure ensemble using a novel similarity metric. *J Biomol Struct Dyn*, **25**, 685-696.
138. Heyne, S., Costa, F., Rose, D. and Backofen, R. (2012) GraphClust: alignment-free structural clustering of local RNA secondary structures. *Bioinformatics*, **28**, i224-232.
139. Sakakibara, Y., Popendorf, K., Ogawa, N., Asai, K. and Sato, K. (2007) Stem kernels for RNA sequence analyses. *J Bioinform Comput Biol*, **5**, 1103-1122.
140. Schattner, P. (2002) Searching for RNA genes using base-composition statistics. *Nucleic Acids Research*, **30**, 2076-2082.
141. Jiang, M., Anderson, J., Gillespie, J. and Mayne, M. (2008) uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. *BMC Bioinformatics*, **9**, 192.
142. Izzo, J.A., Kim, N., Elmetwaly, S. and Schlick, T. (2011) RAG: an update to the RNA-As-Graphs resource. *BMC Bioinformatics*, **12**, 219.
143. Fan, W., Li, J., Ma, S., Tang, N., Wu, Y. and Wu, Y. (2010) Graph Pattern Matching: From Intractable to Polynomial Time. *Proceedings of the VLDB Endowment*, **3**, 264-275.
144. Novikova, I.V., Hennelly, S.P. and Sanbonmatsu, K.Y. (2012) Structural architecture of the human long non-coding RNA, steroid receptor RNA activator. *Nucleic Acids Research*, **40**, 5034-5051.
145. Zhang, B., Yehdego, D.T., Johnson, K.L., Leung, M.Y. and Taufer, M. (2013) Enhancement of accuracy and efficiency for RNA secondary structure prediction by sequence segmentation and MapReduce. *BMC Struct Biol*, **13 Suppl 1**, S3.
146. Lu, Z.J., Gloor, J.W. and Mathews, D.H. (2009) Improved RNA secondary structure prediction by maximizing expected pair accuracy. *RNA*, **15**, 1805-1813.
147. Singhal, A. (2001) Modern Information Retrieval: A Brief Overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, **24**, 35-42.
148. Dice, L.R. (1945) Measures of the Amount of Ecologic Association Between Species. *Ecology*, **26**, 297-302.
149. Sørensen, T. (1948) *A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons.* København, I kommission hos E. Munksgaard, Kongelige Danske Videnskabernes Selskab.
150. Hamming, R. (1950) Error detecting and error correcting codes. *Bell Syst Tech J*, **29**, 147-160.
151. Jaccard, P. (1912) The distribution of the flora in the alpine zone. *New Phytologist*, **11**, 37-50.

152. Rose, P.W., Beran, B., Bi, C., Bluhm, W.F., Dimitropoulos, D., Goodsell, D.S., Prlic, A., Quesada, M., Quinn, G.B., Westbrook, J.D. *et al.* (2011) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Research*, **39**, D392-D401.
153. Yang, H.W., Jossinet, F., Leontis, N., Chen, L., Westbrook, J., Berman, H. and Westhof, E. (2003) Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Research*, **31**, 3450-3460.
154. Brown, J.W. (1999) The Ribonuclease P Database. *Nucleic Acids Research*, **27**, 314-314.
155. Mao, C., Bhardwaj, K., Sharkady, S.M., Fish, R.I., Driscoll, T., Wower, J., Zwieb, C., Sobral, B.W. and Williams, K.P. (2009) Variations on the tmRNA gene. *RNA Biol*, **6**, 355-361.
156. Cannone, J.J., Subramanian, S., Schnare, M.N., Collett, J.R., D'Souza, L.M., Du, Y.S., Feng, B., Lin, N., Madabusi, L.V., Muller, K.M. *et al.* (2002) The Comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*, **3**, 31.
157. Mathews, D.H. (2004) Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, **10**, 1178-1190.
158. Duan, S., Mathews, D.H. and Turner, D.H. (2006) Interpreting oligonucleotide microarray data to determine RNA secondary structure: application to the 3' end of Bombyx mori R2 RNA. *Biochemistry*, **45**, 9819-9832.
159. Harmanci, A.O., Sharma, G. and Mathews, D.H. (2009) Stochastic sampling of the RNA structural alignment space. *Nucleic Acids Research*, **37**, 4063-4075.
160. Samudrala, R. and Levitt, M. (2000) Decoys 'R' Us: a database of incorrect conformations to improve protein structure prediction. *Protein Sci*, **9**, 1399-1401.
161. Simons, K.T., Ruczinski, I., Kooperberg, C., Fox, B.A., Bystroff, C. and Baker, D. (1999) Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins*, **34**, 82-95.
162. Wang, K., Fain, B., Levitt, M. and Samudrala, R. (2004) Improved protein structure selection using decoy-dependent discriminatory functions. *BMC Struct Biol*, **4**, 8.
163. Tsai, J., Bonneau, R., Morozov, A.V., Kuhlman, B., Rohl, C.A. and Baker, D. (2003) An improved protein decoy set for testing energy functions for protein structure prediction. *Proteins*, **53**, 76-87.
164. Samudrala, R. and Moult, J. (1998) An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J Mol Biol*, **275**, 895-916.
165. Handl, J., Knowles, J. and Lovell, S.C. (2009) Artefacts and biases affecting the evaluation of scoring functions on decoy sets for protein structure prediction. *Bioinformatics*, **25**, 1271-1279.

166. Henderson, B.R., Menotti, E., Bonnard, C. and Kühn, L.C. (1994) Optimal sequence and structure of iron-responsive elements. Selection of RNA stem-loops with high affinity for iron regulatory factor. *J Biol Chem*, **269**, 17481-17489.
167. Piccinelli, P. and Samuelsson, T. (2007) Evolution of the iron-responsive element. *RNA*, **13**, 952-966.
168. Diestel, R. (2000) *Graph theory*. Springer-Verlag New York, Electronic Edition.
169. Freeman, L.C. (1979) Centrality in social networks conceptual clarification. *Social networks* **1**, 215-239.
170. Tore Opsahl and Panzarasa, P. (2009) Clustering in Weighted Networks. *Social Networks*, **31**, 155-163.
171. P, S. and A, D.C. (2006) Hairpin RNA: a secondary structure of primary importance. *Cell Mol Life Sci*, **63**, 901-908.
172. Rajaraman, A. and Ullman, J. D. (2011) Mining of Massive Datasets, Cambridge University Press, Cambridge, UK, pp. 1-17.
173. Wikipedia. http://en.wikipedia.org/wiki/Vector_space_model.

VITA

VITA

Jiajie Huang

Department of Biological Sciences, Purdue University

Interdisciplinary Life Science Program (PULSe), Purdue University

Education

B.S., Biological Sciences, 2009, Fudan University, Shanghai, China

M.S., Applied Statistics, 2014, Purdue University, West Lafayette, Indiana

Ph.D., Computational Biology, 2016, Purdue University, West Lafayette, Indiana

Career Interests

Bioinformatician/Biostatistician positions in Thermo Fisher Scientific

PUBLICATIONS

PUBLICATIONS

J. Huang, K. Li, and M.R. Gribskov. “Accurate classification of RNA structures using topological fingerprints.” Submitted.

J. Huang and M.R. Gribskov. “Identification of RNA structural ensembles with pseudoknots using combination of multiple prediction programs.” Submitted.

J. Huang and M.R. Gribskov. “Computational design of decoy RNA structures using a graphical approach.” In preparation.

J. Huang, K. Li, and M.R. Gribskov. “Identification of conserved RNA structural motifs using a subgraph random sampling approach.” Oral Presentation, Great Lakes Bioinformatics Conference, West Lafayette, IN (2015).

J. Huang, K. Li, and M.R. Gribskov. “Identification of conserved RNA structural motifs using a subgraph random sampling approach.” Poster Presentation, Annual International Conference on Research in Computational Molecular Biology (RECOMB), Pittsburgh, PA (2014).

J. Huang. “Graphical approaches for RNA structure matching.” Invited Presentation, Beijing Genomics Institute, Shenzhen, China (2014).

J. Huang, K. Li, and M.R. Gribskov. “Identification and classification of conserved RNA structural motifs using a graph theoretical approach.” Poster Presentation, Annual International Conference on Intelligent Systems for Molecular Biology (ISMB), Long Beach, CA (2012).

J. Huang, K. Li, and M.R. Gribskov. “A graph theoretical approach on RNA functional motif identification.” Poster Presentation, Great Lakes Bioinformatics Conference, Ann Arbor, MI (2012).

X. Zhu, J. Cai, **J. Huang**, X. Jiang, and D. Ren. “The Treatment and Prevention of Mouse Melanoma with an Oral DNA Vaccine Carried by Attenuated Salmonella Typhimurium.” Journal of Immunotherapy 33(5):453-60 (2010).